Unit 4

Population means and variances

Introduction

At the start of Unit 2, data on the lengths (in cm) of a random sample of 100 leaves from an ornamental bush were introduced. A histogram of these data was provided, which is repeated here, in Figure 1(a), for convenience. These data were used to illustrate some considerations in modelling such data in Sections 1 and 2 of Unit 2. In particular, it was emphasised that a *sample* of values such as these is usually collected in order to learn about the *population* of values from which the sample is drawn. It was argued that a theoretical probability distribution for the population of leaf lengths might look something like Figure 12 of Unit 2, which is also repeated here (without a currently irrelevant shaded area), in Figure 1(b), for convenience.

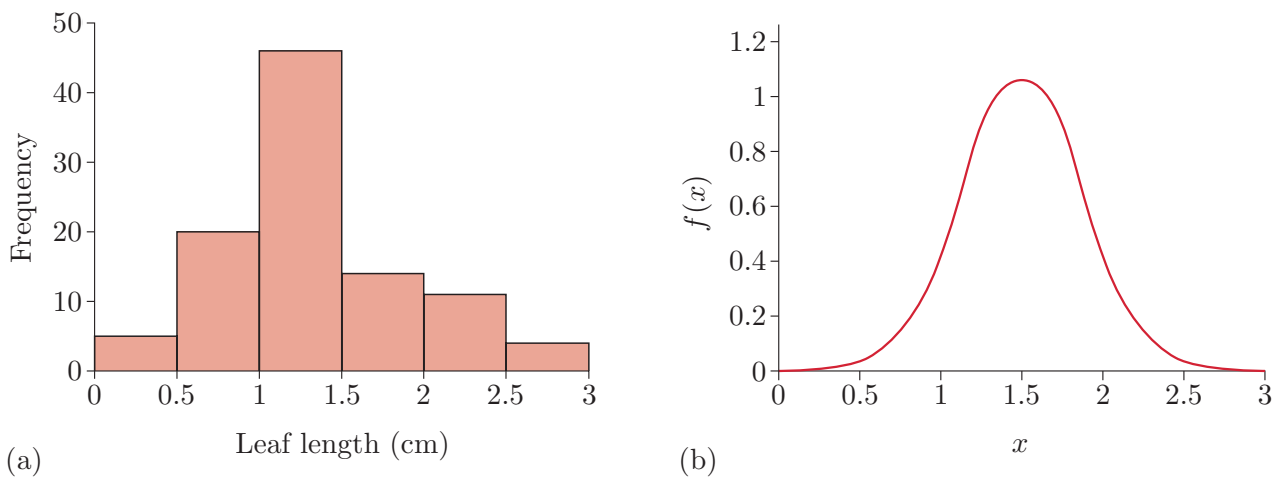


Figure 1 (a) Histogram of leaf lengths; (b) probability density function of a theoretical model for leaf lengths

In Unit 1, you were reminded that it is also often very useful to summarise data like these using a small number of numerical values. Prominent among these are the average of the values in the sample, the sample mean

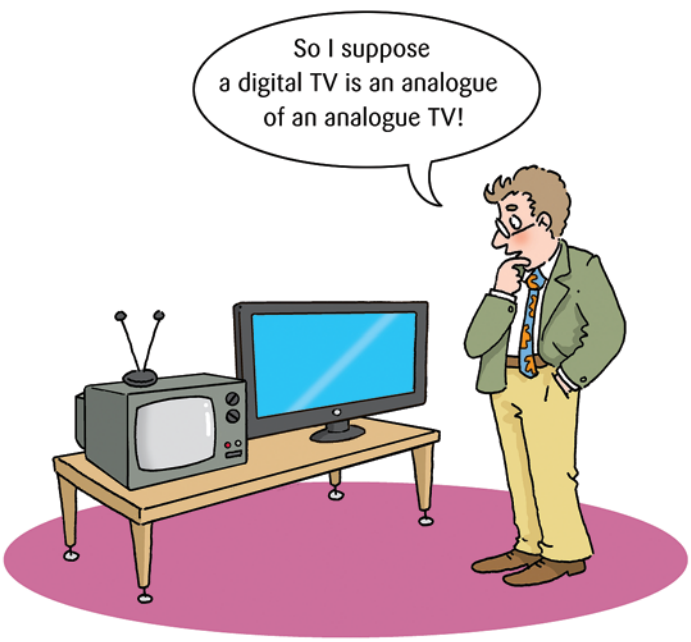
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which is a measure of the location of the sample, and the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

which, along with its square root, s, the sample standard deviation, is a measure of the spread or dispersion of the sample. Here, the data are represented by $x_1, x_2, \ldots, x_n$ where n is the sample size. For the leaf length data, these sample quantities take the numerical values $\overline{x} = 1.276$ cm, $s^2 \simeq 0.319$ cm^2 and $s \simeq 0.564$ cm.

Well, if samples have numerical summaries (like the sample mean, sample variance and sample standard deviation) and if samples give rise to models for the population from which the sample was drawn (like the models you studied in Units 2 and 3), do those models also have numerical summaries that are analogues of the sample ones? The answer is 'yes' and provides the central topic of this unit.



In particular, in Sections 1 to 4, population analogues of the sample mean and sample variance are described; specifically, we look at what is meant by the mean and variance of a probability distribution and at how they are calculated. These quantities are often called the population mean and the population variance. The population mean is considered separately for discrete and continuous distributions in Sections 1 and 2, respectively; similarly, the population variance is considered separately for discrete and continuous distributions in Sections 3 and 4, respectively. Just as the square root of the sample variance is the sample standard deviation, so the square root of the population variance is defined to be the population standard deviation; this also appears from Section 3 onwards. Notice, however, that we concentrate on the population variance rather than the population sample deviation because working with the variance keeps the kinds of calculation performed on probability distributions slightly simpler. (The square root in the standard deviation rather gets in the way!)

It is pretty standard practice to quote means and variances, not standard deviations, of probability distributions.

Finally, in Subsection 5.1, we consider the population means and variances of the distributions of simple, linear, functions of a random variable. A particular application of the work of Subsection 5.1, to simplifying the calculation of population variances, is made in Subsection 5.2.

1 The mean of a discrete distribution

If we assume that a particular probability model is an accurate reflection of the variation in the population, then what, *according to the model*, is the mean of the population? In this section and the next, we look at how a numerical summary of a probability model analogous to the sample mean is defined. This is the **population mean**, or simply the **mean**, of the probability distribution. In particular, in this section, we focus on the discrete case: in Subsection 1.1, the mean of a discrete distribution is defined; and in Subsection 1.2, formulas are obtained for the means of members of the five families of discrete probability distributions that you have met so far in this module. The mean of a continuous probability distribution will be discussed in Section 2.

1.1 The population mean in the discrete case

The two ideas needed to develop a formula for the mean of a discrete probability distribution are the way a sample mean is calculated and the definition of a probability. We begin by considering the mean score obtained when an unbiased six-sided die is rolled.

Example 1 *Rolls of an unbiased six-sided die*

The outcomes of 30 rolls of an unbiased die are given in Table 1.

Table 1 Outcomes of 30 rolls of an unbiased die

4	3	2	5	6	3	2	6	4	3	6	4	3	2	5
3	1	5	2	1	4	3	1	2	1	1	2	5	2	5

The sample mean for the 30 rolls may be calculated by finding the sample total and dividing it by the sample size:

$$\overline{x} = \frac{4+3+2+\cdots+2+5}{30} = \frac{96}{30} = 3.2.$$

Alternatively, we could start by summarising the original data in the form of a frequency table. This is shown in Table 2.

Table 2 Frequency table for the 30 rolls of a die in Table 1

Outcome (x)	1	2	3	4	5	6
Frequency (f_x)	5	7	6	4	5	3

Table 2 is obtained by noting that there are five 1s in Table 1, seven 2s, and so on. The five 1s give a total of $1 \times 5 = 5$, the seven 2s a total of $2 \times 7 = 14$, etc. So the sample total is

$$(1 \times 5) + (2 \times 7) + (3 \times 6) + (4 \times 4) + (5 \times 5) + (6 \times 3)$$
$$= 5 + 14 + 18 + 16 + 25 + 18 = 96.$$

Above: trying to speed up the process!

Then the sample mean can be computed by dividing this total by the sample size:

$$\overline{x} = \frac{96}{30} = 3.2.$$

As expected, this method leads to the same result.

This example shows that, given a sample of data from a discrete distribution, there are two equivalent ways of calculating the sample mean. If $x_1, x_2, \ldots, x_{30}$ denote the values in the sample in Example 1, then in the first method we obtain the sample mean by adding all the values together and dividing by 30. In this case we are using the familiar definition of the sample mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1}$$

In the second approach, we first count how many of each particular outcome there are in the sample in order to obtain a frequency table. If we denote the number of occurrences in the sample of outcome x by f_x (for instance, $f_2 = 7$ in Example 1), then for each x the contribution made to the total by the outcome x is $x \times f_x$ (for example, $2 \times 7 = 14$). Adding up these contributions and then dividing by n (the sample size, which is equal to 30 in Example 1), we obtain the sample mean in the form

$$\overline{x} = \frac{1}{n}\sum_{x} x f_x. \tag{2}$$

Here the sum is over all possible outcomes x ($1, 2, \ldots, 6$ for the die). The two equations (1) and (2) for $\overline{x}$ give the same answer.

It is helpful to rewrite Equation (2) in the form

$$\overline{x} = \sum_{x} x \frac{f_x}{n},$$

where f_x/n is the sample relative frequency of outcome x. Now recall the definition of a probability from Unit 2: $p(x)$, the probability that the observation x occurs, is the limiting value of the sample relative frequency of the observation as the sample size n gets larger and larger; that is, it is the limiting value of f_x/n.

So if we were to keep increasing the sample size n, then the sample relative frequency of the observation x would tend to get closer and closer to $p(x)$, and the sample mean $\overline{x}$ would approach $\sum_x x\,p(x)$.

This is another instance of the 'settling down' phenomenon that you investigated in Unit 2, both for individual proportions tending to probabilities and for whole sets of sample relative frequencies tending to probability mass functions.

Example 2 *The mean score when an unbiased die is rolled*

For a die that is assumed to be unbiased, each of the six possible outcomes $x = 1, 2, \ldots, 6$ occurs with probability $p(x) = \frac{1}{6}$. So in a very large number of rolls of an unbiased die, the mean score should be approximately

$$\sum_{x=1}^{6} x\,p(x) = \left(1 \times \tfrac{1}{6}\right) + \left(2 \times \tfrac{1}{6}\right) + \left(3 \times \tfrac{1}{6}\right) + \left(4 \times \tfrac{1}{6}\right) + \left(5 \times \tfrac{1}{6}\right) + \left(6 \times \tfrac{1}{6}\right)$$
$$= \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5.$$

The definition of the population mean for a discrete random variable X (or for a discrete probability distribution) with probability mass function $p(x)$ is based on the idea illustrated above for the score on an unbiased die. When referring to a population mean, the terms 'population mean' and 'mean' are both commonly used. We also sometimes refer to the mean of a random variable and on other occasions to the mean of a probability distribution. All of these expressions are equivalent: they all refer (in the discrete case) to the numerical summary defined in the following box.

The population mean of a discrete distribution

For a discrete random variable X with probability mass function $p(x)$, the **(population) mean** of X (or of the probability distribution of X) is denoted by μ and is given by

$$\mu = \sum_{x} x\,p(x), \tag{3}$$

where the sum is over all possible values of X, that is, over all values in the range of X.

μ is the Greek lower-case letter mu, pronounced 'mew'.

From this definition it follows that if X is a random variable denoting the score obtained when an unbiased die is rolled, then (from Example 2) the mean of the random variable X is 3.5.

Activity 1 *The mean score on a biased die*

In Activity 8(b) of Unit 2, a particular biased die was introduced. You saw that, for a die with the face that should show two spots replaced by five spots, the probability distribution for the outcome of a single roll is as given in Table 3.

Find the mean score obtained when such a die is rolled. How does it compare with the mean score obtained when an unbiased die is rolled?

Table 3 The p.m.f. for a die with two faces showing five

w	1	3	4	5	6
$p(w)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$

A point to note is that it is not necessary for the mean of a random variable to be a value in the range of the random variable: 3.5 is the mean but it is not a possible outcome when an unbiased die is rolled. As another example, explored in Activity 2, the mean family size in the UK is not an integer.

Activity 2 *The mean family size in the UK*

The terms 'dependent children' and 'families' have precise definitions which need not concern you now.

This activity concerns the numbers of dependent children in families in the UK with at least one dependent child. According to the Office for National Statistics, in 2012, the population of such families consisted of 3.7 million families with one child, 3 million families with two children, and 1.1 million families with three or more children. As an approximation, take 'three or more children' to be exactly 'three children'. The data are then summarised in the following frequency table.

Table 4 Numbers of children in UK families in 2012

Number of children (x)	1	2	3
Frequency (f_x)	3 700 000	3 000 000	1 100 000

(a) Find the mean number of children in UK families with at least one dependent child in 2012.

(b) What do you think has been the effect of replacing 'three or more children' by 'three children'?

An alternative terminology and its notation are widely used for the mean of a random variable. They are given in the box below. Notice that this terminology and notation apply to means of all random variables and distributions, both discrete and continuous.

More terminology and notation for means

The (population) mean of a random variable X is also called the **expected value** of X or the **expectation** of X. The corresponding alternative notation to μ is $E(X)$, that is,

$$\mu = E(X).$$

$E(X)$ is usually read as 'the expected value of X' or simply as 'E of X'.

Example 3 *Using the E notation in the discrete case*

If X denotes the score obtained when an unbiased die is rolled (see Example 2), then the expected value of X is

$$E(X) = 3.5.$$

If the random variable W denotes the score obtained when a die with its two-spot face replaced by a five-spot face is rolled (see Activity 1), then

the expected value of W is

$$E(W) = 4.$$

Finally, if Y is the number of dependent children in a family with at least one dependent child in the UK in 2012 (see Activity 2), then the expected number of children in such a family is

$$E(Y) \simeq 1.67.$$

Occasionally, the phrase 'expected value' or 'expected number' is more natural than 'mean'. But notice that 'the value you would expect to get' is not a valid interpretation: you could not actually have 1.67 children in a family, because 1.67 is not an integer.

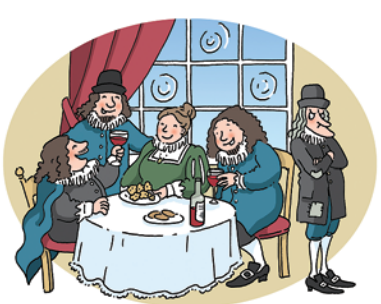
The notation μ for the mean of a random variable X is sometimes modified to include the subscript X, that is, μ_X. This notation is particularly useful in situations involving more than one random variable: the means of the different random variables cannot then be confused. However, in this module a subscript will not usually be included unless it is necessary to avoid ambiguity.

1.2 Families of distributions: the mean

So far, the examples and activities have involved probability distributions where the probability mass function is specified exactly. In Unit 3, the idea of indexing distributions by some quantity or quantities called parameter(s) of the model was introduced. These parameters are sometimes known quantities but are usually unknown quantities. In this subsection, for each of the five families of discrete probability distributions so far discussed, expressions for the mean which involve the parameter(s) of the model are established. These families of distributions are the Bernoulli, binomial, geometric, Poisson and discrete uniform distributions.

The mean of a Bernoulli distribution

The first family of discrete distributions to which you were introduced in Unit 3 is the Bernoulli family. Each distribution in this family allows only the two possible outcomes 0 or 1, and the probability mass function of a Bernoulli distribution is $p(1) = p$, $p(0) = 1 - p$. Here, p is the indexing parameter.



Most of the Bernoullis were generous, but one of them was mean

Activity 3 *The mean of a Bernoulli distribution*

Use Equation (3) to find an expression for the population mean of any Bernoulli distribution in terms of p.

In Activity 3, you obtained the important result given in the following box.

The mean of a Bernoulli distribution

If the random variable X has a Bernoulli distribution with parameter p, $X \sim \text{Bernoulli}(p)$, then

$$\mu = E(X) = p. \tag{4}$$

Notice that because the range of the Bernoulli distribution is the values 0 and 1, and the mean of the Bernoulli distribution is p, which satisfies $0 < p < 1$, the mean of the Bernoulli distribution is never a member of the range of X. In repeated sampling from a Bernoulli distribution, we sometimes get '1' and we sometimes get '0', so the average of the values that we get is something in between.

Activity 4 *Means of Bernoulli distributions*

(a) What is the mean score resulting from a toss of a fair coin, if we score 1 for heads and 0 for tails?

(b) Suppose that a random variable X is defined to take the value 1 when an unbiased die shows a 3 or a 6, and 0 otherwise. The probability distribution for X is given by

$$P(X = 0) = \tfrac{2}{3}, \qquad P(X = 1) = \tfrac{1}{3}.$$

What is the mean of the distribution of X?

The mean of a binomial distribution

The second family of discrete probability distributions introduced in Unit 3 is the binomial family. If the random variable X has a binomial distribution with parameters n and p, $X \sim B(n, p)$, then the probability mass function of X is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

So, using the definition in Equation (3), the mean of X (or the expected value of X) is given by

$$\mu = E(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}.$$

In fact, this calculation is not as difficult as it looks, and a certain amount of algebraic manipulation would give us the answer we need. However, almost no algebra is necessary if we think about what the binomial random variable X represents. It is the number of successes in a sequence of n independent trials, where the probability of success at each trial is p. So the problem posed is this: what is the expected number of successes in such a sequence of n trials?

The number of successes in a single trial has a Bernoulli distribution with parameter p, so the expected number of successes in a single trial is p. This suggests that the expected number of successes in n trials should be $n \times p$. This result can be established formally as follows.

A general result for means states that the mean of a sum of random variables is equal to the sum of the means of the random variables:

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n). \tag{5}$$

This result will not be proved in this module. It is one manifestation of the general term 'linearity of expectation'.

A binomial random variable, $B(n, p)$, is the sum of n Bernoulli(p) random variables, so its mean is

$$\underbrace{p + p + \cdots + p}_{n \text{ terms}} = np.$$

That is, the mean of a binomial random variable X with parameters n and p is given by the product np.

The mean of a binomial distribution

If the random variable X has a binomial distribution with parameters n and p, $X \sim B(n, p)$, then

$$\mu = E(X) = np. \tag{6}$$

Again, the mean of a binomial distribution may be some value not in the range of X. For instance, if $n = 100$ and $p = \frac{1}{3}$, then the mean is $np = 100 \times \frac{1}{3} = 33\frac{1}{3}$, which is not an integer. The mean or expectation is a statement about the 'long-term' average number of successes in sequences of Bernoulli trials, and thus need not be an integer.

Example 4 *The mean of a binomial random variable: two methods of calculation and one of interpretation*

If X has a binomial distribution with parameters 4 and 0.4, $X \sim B(4, 0.4)$, then its probability mass function is given by

$$p(x) = \binom{4}{x} (0.4)^x (0.6)^{4-x}, \quad x = 0, 1, 2, 3, 4.$$

The individual probabilities are as follows.

Table 5

x	0	1	2	3	4
$p(x)$	0.1296	0.3456	0.3456	0.1536	0.0256

Then, using the definition of the mean of a discrete random variable, we obtain

$$\begin{aligned} E(X) &= \sum_{x=0}^{4} x\, p(x) = (0 \times 0.1296) + (1 \times 0.3456) + \cdots + (4 \times 0.0256) \\ &= 0 + 0.3456 + 0.6912 + 0.4608 + 0.1024 = 1.6. \end{aligned}$$

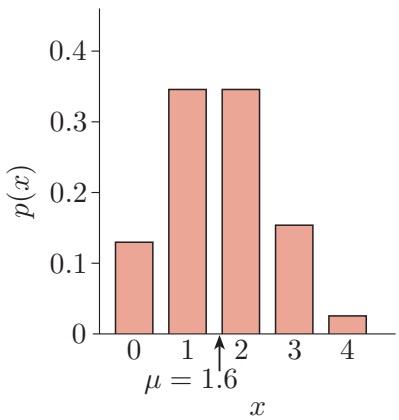


Figure 2 The p.m.f. and mean of X when $X \sim B(4, 0.4)$

This result is obtained much more easily using Equation (6):

$$E(X) = np = 4 \times 0.4 = 1.6.$$

This mean is represented graphically in Figure 2. The point $x = 1.6$ (the mean of the distribution) is shown using an arrow on a sketch of the probability mass function of X.

The mean has the following physical interpretation (idealised), which you might or might not find helpful. Imagine weights of mass 0.1296, 0.3456, 0.3456, 0.1536 and 0.0256 units placed at equal intervals on a thin plank of zero mass. If the plank is represented by the horizontal axis in Figure 2, then it will balance at the point indicated by the arrowhead – in this case, at a point just to the right of the midpoint between the two largest weights.

Any defectives here?

Activity 5 *Means of binomial distributions*

Each of the following situations involves a random variable with a binomial distribution. Use Equation (6) to find the mean in each case.

(a) The probability that an item from a production line is defective is 0.01. What is the expected number of defective items in a sample of 100 items taken from the production line?

(b) The probability that an archer hits the centre of the target with each arrow that she shoots is 3/4. Find the mean number of arrows that hit the centre of the target in 10 shots.

(c) The probability that a tennis player wins each match he plays against a friend is 0.7. Find his expected number of wins if he plays five matches.

The mean of a geometric distribution

The third family of discrete distributions introduced in Unit 3 is the geometric family. The geometric distribution is a model for the number of trials up to and including the first success in a sequence of independent Bernoulli trials. If the probability of success in each trial is p, then how many trials, on average, are required to obtain a success? That is, what is the mean of a geometric distribution with parameter p?

Activity 6 *Guessing the mean*

Without doing any calculations, try to answer the following questions. (Just jot down your first reaction.)

(a) The probability that a fair coin shows heads when it is tossed is $\frac{1}{2}$. How many times, on average, do you think a coin would need to be tossed to come up heads?

(b) The probability that a die with its two-spot face replaced by a second five-spot face comes to rest showing '5' when it is rolled is $\frac{1}{3}$. On average, how many times do you think such a die would need to be rolled to show a 5?

(c) Approximately one car in six on British roads is grey. If you stand by the side of the road and start counting, how many cars, on average, do you think you will have to count to record your first grey one?

Plenty of grey cars here

Each of the situations in Activity 6 involves a random variable with a geometric distribution. A formula for the mean of a geometric distribution with parameter p is provided below. In Activity 7, you will be asked to use this formula to find the mean of each of the random variables described in Activity 6. When you have done so, compare your answers to Activity 6 with those obtained using the formula.

From Unit 3, the p.m.f. of a geometric distribution with parameter p is

$$p(x) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots .$$

In Screencast 4.1, the mean of a geometric random variable X with parameter p is derived using this p.m.f. This screencast is, however, optional: you will not therefore be expected to reproduce the algebraic details given there.

Screencast 4.1 Derivation of the mean of a geometric distribution (optional)

The result obtained in Screencast 4.1 is given in the following box.

The mean of a geometric distribution

If the random variable X has a geometric distribution with parameter p, then

$$\mu = E(X) = \frac{1}{p}. \tag{7}$$

Notice that this means that the less likely an event is to occur, the longer you should expect to wait for it to happen.

Activity 7 *Means of geometric distributions*

Each of the situations described in Activity 6 involves a random variable with a geometric distribution. Use Equation (7) to find the mean for each of these random variables. Compare these values with the answers you gave in Activity 6.

Given appropriate data, the question arises as to how to choose a value for the parameter p when fitting a geometric model to the data. The formula for the mean in Equation (7) suggests a possible approach. We would like the data and the model to have similar means, and rearranging Equation (7) gives

$$p = \frac{1}{\mu}.$$

So a common sense estimate for p is the reciprocal of the sample mean: 1/(sample mean). This formula is, in fact, the one that is obtained using a formal method of estimation that is discussed later in this module.

You will use this formula in part (c) of the following activity which considers various versions of the problem of estimating p from data relating to Bernoulli trials.

Activity 8 *Estimating parameter values*

The probability that a darts player hits the bull's-eye with each dart that she throws is equal to p. This probability is constant from dart to dart, so if she throws several darts, the throws may be regarded as a sequence of independent Bernoulli trials.

(a) In 50 attempts, the player hits the bull's-eye on 12 occasions. What would you estimate the value of p to be?

(b) The numbers of bull's-eyes obtained in each of ten sequences of 20 throws are as follows.

4 10 2 6 5 5 9 3 6 4

Use these data to obtain an estimate for the value of p.

(c) The number of throws needed to obtain a bull's-eye was recorded on 30 occasions. The data are given in Table 6. Use these data to obtain an estimate of p.

Table 6 The number of throws needed to obtain a bull's-eye

Number	1	2	3	4	5	6	7	10	12	15
Frequency	8	5	4	3	3	2	2	1	1	1

The mean of a Poisson distribution

In Unit 3, we also introduced the Poisson distribution. This is a model for counts, with range $\{0, 1, 2, \ldots\}$. It has p.m.f.

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

with its parameter λ being positive. So if $X \sim \text{Poisson}(\lambda)$, then the mean

of X is given by

$$E(X) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}.$$

The first term in this sum is zero, so

$$E(X) = \sum_{x=1}^{\infty} \frac{xe^{-\lambda}\lambda^x}{x!}.$$

For each term in this sum, the x in the numerator and the x in the $x! = x \times (x-1)!$ term in the denominator can be cancelled, so

$$E(X) = \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!}.$$

Don't worry too much about the algebraic details to follow. There is a screencast offering further help below.

Taking a factor $\lambda e^{-\lambda}$ outside the summation (because it consists of constants not dependent on x) gives

$$\begin{aligned} E(X) &= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda} \left(\frac{1}{0!} + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots \right) \\ &= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \lambda \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}. \end{aligned}$$

Now, the summation is actually $\sum_{x=0}^{\infty} p(x)$, that is, the sum of all the values of the Poisson p.m.f. over its range. This summation therefore has value 1. Hence

$$E(X) = \lambda \times 1 = \lambda.$$

The following screencast offers a little more help with the above calculation.

Screencast 4.2 Obtaining the Poisson mean

The parameter of the Poisson distribution has therefore turned out to be its mean. This result is highlighted in the following box.

The mean of a Poisson distribution

If the random variable X has a Poisson distribution with parameter λ, then

$$\mu = E(X) = \lambda. \tag{8}$$

Activity 9 *Means of Poisson distributions*

In Section 4 of Unit 3, the following random variables were argued to be appropriately modelled by Poisson distributions. In each case, you are asked to give the value of the mean of the random variable.

(a) If X is the number of claims on a motor insurance policy over a 5-year period and follows the Poisson distribution with parameter $\lambda = 0.5$, what is $E(X)$?

(b) If Y is the number of yeast cells found in a randomly chosen small square on a microscope slide and follows the Poisson distribution with $\lambda = 0.6825$, what is $E(Y)$?

The mean of a discrete uniform distribution

The discrete uniform distribution on the integers $m, m+1, \ldots, n$ was introduced in Subsection 5.1 of Unit 3. An important special case of this, the case with $m = 1$, is the discrete uniform distribution on the first n integers which has probability mass function

$$p(x) = \frac{1}{n}, \quad x = 1, 2, \ldots, n. \tag{9}$$

For example, the score on an unbiased six-sided die follows the distribution with p.m.f. given by Equation (9) when $n = 6$. In Example 2, it was shown that the mean score on an unbiased six-sided die is 3.5. In Activity 10, you will consider other similar examples from which you are asked to suggest a general formula for the mean of the discrete uniform distribution on the range $1, 2, \ldots, n$.

Activity 10 *Means of discrete uniform distributions on the first n integers*

(a) Find the mean score obtained when an unbiased tetrahedral die (that is, a die with four faces) is rolled.

(b) A positive digit, that is, an integer between 1 and 9 inclusive, is chosen at random (that is, so that no number is more likely than any other to be chosen). Find the expected value of the number chosen.

(c) Given the examples above, can you suggest a formula for the mean of the discrete uniform distribution on the first n integers?

Activity 11 *The mean of the discrete uniform distribution on the first n integers*

You might already know the mathematical result that the sum of the first n integers is equal to $n(n+1)/2$, that is,

$$\sum_{x=1}^{n} x = \frac{n(n+1)}{2}. \tag{10}$$

Use this result in conjunction with Equation (3) to find an expression for the population mean of the discrete uniform distribution with p.m.f. given by Equation (9), in terms of n.

Activity 12 *Applying the result of Activity 11 to Activity 10*

Each of the situations described in parts (a) and (b) of Activity 10 involves a random variable with the discrete uniform distribution on the first n integers. Use the formula you obtained in Activity 11 to confirm the values of the mean that you obtained in Activity 10 for each of these random variables.

More generally, the discrete uniform distribution with parameters m and n has probability mass function

$$p(x) = \frac{1}{n-m+1}, \quad x = m, m+1, \ldots, n.$$

Now,

$$E(X) = \sum_{x=m}^{n} \frac{x}{n-m+1} = \frac{1}{n-m+1} \sum_{x=m}^{n} x.$$

The following argument is given for the case that $m = 1, 2, \ldots$; the result in the box below holds for any $m = \ldots, -2, -1, 0, 1, 2, \ldots$. The summation term is $m + (m+1) + \cdots + n$ which is the sum of the first n integers minus the sum of the first $m-1$ integers. Therefore, using Equation (10),

$$\begin{aligned} E(X) &= \frac{1}{n-m+1}\left(\sum_{x=1}^{n} x - \sum_{x=1}^{m-1} x\right) \\ &= \frac{1}{n-m+1}\left(\frac{n(n+1)}{2} - \frac{(m-1)m}{2}\right) \\ &= \frac{n^2+n-m^2+m}{2(n-m+1)}. \end{aligned}$$

Using the fact that $n^2 + n - m^2 + m = (n-m+1)(n+m)$, this becomes

$$E(X) = \frac{(n-m+1)(n+m)}{2(n-m+1)} = \frac{n+m}{2}.$$

The mean of a discrete uniform distribution

If the random variable X has a discrete uniform distribution on $m, m+1, \ldots, n$, then the mean of X is given by

$$E(X) = \frac{n+m}{2}. \tag{11}$$

The mean of a random variable following the discrete uniform distribution on all the integers from m to n is therefore the average of the values m and n, and hence falls midway between the two ends of the range. Also, Equation (11) reduces to $E(X) = (n+1)/2$ when $m = 1$, the result you found in Activity 11.

Activity 13 *The mean of the discrete uniform distribution on the digits including zero*

In part (b) of Activity 10, the mean of the discrete uniform distribution on positive digits, that is, on $1, 2, \ldots, 9$, was found to be 5. Find the mean of the discrete uniform distribution on the digits including zero, that is, on $0, 1, \ldots, 9$. What is the effect on the mean of adding zero to the list of possible outcomes?

Exercises on Section 1

Exercise 1 *Practice at finding means*

The probability distributions of the random variables X and Y are given in Tables 7 and 8.

Table 7 The p.m.f. of X

x	2	3	4	5
$p(x)$	0.1	0.2	0.3	0.4

Table 8 The p.m.f. of Y

y	0	1	2	3	4
$p(y)$	0.4	0.2	0.1	0.1	0.2

(a) Calculate the expected value of X.

(b) Find the mean of the random variable Y.

Exercise 2 *The mean number of people infected*

An important field that uses statistics a lot is epidemiology: the World Health Organization defines epidemiology as 'the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems'. Many different models have been developed for the transmission of infectious diseases, a few of them simple, most of them rather complicated. In small communities (for instance, families and schools), one variable of interest is the total number of people who contract a disease, given that initially one member of the community becomes infected. In a family of 4, say, that number could be 1, 2, 3 or 4. This number is a random variable because epidemic dynamics are, to a great extent, a matter of chance – whether or not a child catches her brother's cold, for instance, is not a predetermined event.

One model for the spread of a particular disease within a particular family of six people gave the probability distribution in Table 9 for Y, the number who eventually suffer from the disease.

Table 9 The number infected in a family of 6

y	1	2	3	4	5	6
$p(y)$	$\frac{3}{90}$	$\frac{8}{90}$	$\frac{15}{90}$	$\frac{20}{90}$	$\frac{24}{90}$	$\frac{20}{90}$

Find the mean of the distribution of Y.

Exercise 3 *Finding means for standard distributions*

The probability that an archer hits the centre of the target with each arrow she shoots is 0.8.

(a) Suppose that a random variable X is defined to take the value 1 when the archer hits the centre of the target and 0 when she misses. What is the mean of the random variable X?

(b) The archer shoots seven arrows. What is the expected number of arrows that hit the centre of the target?

(c) The archer shoots arrows until she misses the centre of the target. What is the expected value of the number of arrows she shoots?

2 The mean of a continuous distribution

The variation that might be observed in measurements on a discrete random variable is expressed through its probability mass function, and you have seen in Section 1 how to use the p.m.f. to calculate the mean or expected value of a discrete random variable. Similarly, variation observed in measurements on a continuous random variable may be expressed by writing down a probability density function. If this density function is a good model, then we should be able to use it not just for statements about the probabilities of different measurements but also, as in the case of discrete random variables, to provide information about the long-term average in repeated measurements.

So how, in general, is the mean of a continuous random variable calculated?

Equation (3) says that for a discrete random variable X with probability mass function $p(x)$, the mean of X is given by the formula

$$\mu = E(X) = \sum_x x\,p(x),$$

where the summation is taken over all values in the range of X. This represents an average of the different values that X may take, weighted according to their chance of occurrence. The definition of the mean of a continuous random variable, or equivalently of a continuous distribution, is analogous to that of a discrete random variable: the p.d.f. replaces the p.m.f., and integration replaces summation.

The mean of a continuous distribution

For a continuous random variable X with probability density function $f(x)$, the **(population) mean of X** or the **expected value of X** is given by

$$\mu = E(X) = \int x\, f(x)\, dx, \tag{12}$$

where the integral is taken over all values x in the range of X.

As you can see, the technique of integration is required for the calculation of the mean of a continuous random variable, as it was for the calculation of probabilities and of the cumulative distribution function for continuous random variables in Unit 2. The differences here are that you must remember to include an extra 'x' in the integrand (the quantity that you are integrating) and to perform the integration over the entire range of X.

Example 5 *The mean journey time*

He doesn't have to do this every morning: in 2014, commuters in Perth, Australia, freed a man who was trapped by lifting a train

Here's a first example of calculating the mean of a continuous distribution. In Activity 17 of Unit 2, a man's journey to work (in minutes) was represented by a random variable with probability density function

$$f(x) = \frac{1}{5} - \frac{x}{250}$$

on the range $20 < x < 30$. What is the average journey time for this man according to this model?

Applying Equation (12), we find that

$$\begin{aligned} E(X) &= \int_{20}^{30} x\, f(x)\, dx = \int_{20}^{30} x\left(\frac{1}{5} - \frac{x}{250}\right) dx \\ &= \int_{20}^{30} \left(\frac{x}{5} - \frac{x^2}{250}\right) dx = \left[\frac{x^2}{10} - \frac{x^3}{750}\right]_{20}^{30} \\ &= \frac{900}{10} - \frac{27\,000}{750} - \left(\frac{400}{10} - \frac{8000}{750}\right) \\ &= 90 - 36 - \left(40 - \frac{32}{3}\right) = \frac{74}{3} = 24\tfrac{2}{3}. \end{aligned}$$

So, according to the model, the average journey time for this man is $24\frac{2}{3}$ minutes (or 24 minutes and 40 seconds).

In the following activities, you are asked to find the means of some other continuous distributions introduced in Unit 2.

Activity 14 *Means of two continuous distributions*

(a) In Example 19 of Unit 2, you considered the distribution with p.d.f.

$$f(x) = 3x^2, \quad 0 < x < 1.$$

Find the mean of this distribution.

(b) In Example 20 of Unit 2, you considered the distribution with p.d.f.

$$f(x) = 0.6x^2 + 0.2x - 0.7, \quad 1 < x < 2.$$

Find the mean of this distribution.

Activity 15 *The mean length of brown trout fry*

In Exercise 10 of Unit 2, the following distribution was used for the lengths (in cm) of brown trout fry in a hatchery pond:

$$f(x) = \frac{1}{30}(10x - x^2 - 14), \quad 3 < x < 6.$$

Find the mean length of brown trout fry according to this model.

If you are still unsure of the procedure for calculating the mean of a continuous distribution, the following screencast might help.

Screencast 4.3 Obtaining the mean of a continuous distribution

As in the case of the mean of a discrete random variable, the value of the expectation $\mu = E(X)$ of a continuous random variable X has a physical interpretation. It is the point about which a physical model of the area under the p.d.f. would balance.

Example 6 *Means and balances*

In Activity 17 of Unit 2, and Example 5 above, the probability density function shown in Figure 3 (overleaf) was used as a model for the time taken by a man's journey to work (in minutes). Also shown in the figure is the point at which a quadrilateral bounded by the p.d.f. and the axes would balance if the quadrilateral were to be balanced on its bottom edge on a rod represented by the arrow. This point occurs at the value $24\frac{2}{3}$ which, as was shown in Example 5, is the mean of the distribution.

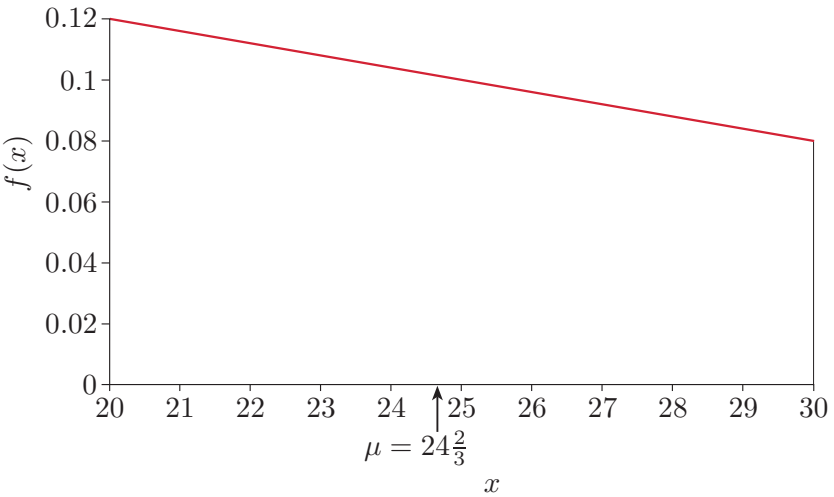


Figure 3 The p.d.f., the completed quadrilateral, and the mean $\mu = 24\frac{2}{3}$

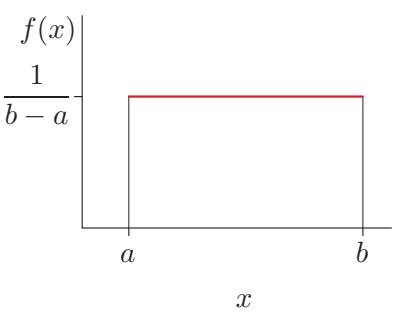


Figure 4 The p.d.f. of $X \sim U(a, b)$

The mean of a continuous uniform distribution

In Subsection 5.2 of Unit 3, a continuous distribution was introduced to reflect the notion of a random variable, X, taking no preferred value between a and b ($b > a$). This continuous uniform distribution, written $U(a, b)$, was seen to have the probability density function

$$f(x) = \frac{1}{b-a}, \quad a < x < b,$$

which is shown in Figure 4.

Activity 16 *The mean of the continuous uniform distribution*

(a) Without doing any calculations, what do you think the mean of the continuous uniform distribution might be?

(b) Now use integration to obtain a formula for the mean of the continuous uniform distribution.

Activity 17 *Means of continuous uniform distributions*

(a) In Activity 26 of Unit 3, it was found that the position, X, of a fault in a cable of length 40 metres might be modelled by a continuous uniform distribution on $(0, 40)$. What is the mean position of such a fault?

(b) What is the mean of a random variable Y which follows the standard continuous uniform distribution with $a = 0$, $b = 1$?

The result on the mean of a continuous uniform distribution is emphasised in the following box.

The mean of a continuous uniform distribution

If the random variable X has a continuous uniform distribution on (a, b), then the mean of X is given by

$$E(X) = \frac{1}{2}(a + b). \tag{13}$$

Exercises on Section 2

Exercise 4 *Practice at finding means*

Find the means associated with each of the distributions whose probability density functions are given below.

(a) $f(x) = 4x^3, \quad 0 < x < 1.$

(b) $f(x) = 3(1 - x)^2, \quad 0 < x < 1.$

(c) $f(x) = 3(x - 1)^2, \quad 1 < x < 2.$

Exercise 5 *The mean of a triangular distribution*

A family of 'triangular' distributions, indexed by parameter $b > 0$, has p.d.f. of the form

$$f(x) = \frac{2(b - x)}{b^2}, \quad 0 < x < b.$$

Find the mean of this triangular distribution (in terms of b).

Exercise 6 *The mean bulldozer return time*

In Activity 25 of Unit 2, a model was introduced for the return times, X, in minutes, of a bulldozer when carrying out a particular earthmoving task. The model has p.d.f.

$$f(x) = \frac{15}{16\sqrt{2}}\sqrt{x}(2 - x), \quad 0 < x < 2.$$

What is the mean bulldozer return time according to this model?

A bull-dozer with a different return time

3 The variance of a discrete distribution

In Sections 1 and 2, the idea of a sample mean was extended to the mean of a theoretical model for the variation observed in data. The mean (or expected value) of a random variable X was denoted by μ or $E(X)$. In this

section, a measure of dispersion for a population, analogous to the sample variance, is defined. Notice again that, whereas when summarising data it is common to find the sample standard deviation, when summarising a probability model it is much more common to quote the variance. The reason for this is simply convenience: many results involving variances of random variables are algebraically simpler than corresponding results involving standard deviations.

The following example illustrates a typical context in which knowledge of a population variance is essential to answering a scientific question.

Example 7 *Measuring intelligence*

A psychologist assessing intellectual ability decides to use the Wechsler Adult Intelligence Scale to measure an individual's intelligence quotient, commonly known as IQ. She finds that one subject has a score of 110. This is above the population mean of 100. But how far above the mean is it? Should she expect many people to score as high as this, or is the difference of ten points a large difference? To answer this question, she needs to know something about the spread or dispersion of IQ scores in the population, and she might, for example, choose to measure this spread using the population analogue of the sample standard deviation or the population analogue of the sample variance.

In Subsection 3.1, the (population) variance of a discrete distribution is defined; and in Subsection 3.2, formulas are obtained for the variances of members of the five families of discrete probability distributions that you have met so far. The (population) variance of a continuous probability distribution will be discussed in Section 4.

3.1 The population variance in the discrete case

In Unit 1, the sample variance was defined to be

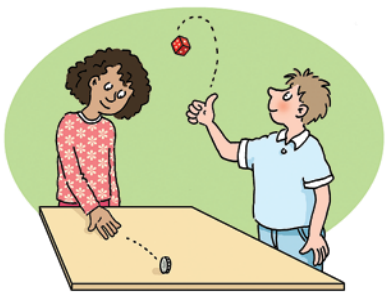
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

This measure of spread gives the average squared deviation of each item in the sample from the sample mean, with the small distinction that the average is obtained through division by $n-1$ rather than by n. The analogous measure for a probability model is the expected squared deviation of a random variable X from the mean of X. This may be written using the 'expectation' notation $E(\text{random variable})$ as

$$E\big[(X-\mu)^2\big].$$

We require not simply 'the expected value of X', which is μ, but also 'the expected value of a function of X', that function being $(X-\mu)^2$, which is itself a random variable. We therefore need to calculate the value of the function $(x-\mu)^2$ for each value x in the range of X, and then average the values of $(x-\mu)^2$, weighting them according to their chance of occurrence.

Example 8 *Rolls of an unbiased die*



In Example 1, we looked at the results of 30 rolls of an unbiased die. The sample mean was found to be $\overline{x} = 3.2$, and the data were given in frequency table form in Table 2. For this sample, the sum of squared deviations from the mean is given by

$$\begin{aligned}&\sum_{i=1}^{30}(x_i - \overline{x})^2\\ &= (1-3.2)^2 + (1-3.2)^2 + (1-3.2)^2 + (1-3.2)^2 + (1-3.2)^2\\ &\quad + (2-3.2)^2 + (2-3.2)^2 + (2-3.2)^2 + (2-3.2)^2 + (2-3.2)^2\\ &\quad + (2-3.2)^2 + (2-3.2)^2 + \cdots + (6-3.2)^2 + (6-3.2)^2 + (6-3.2)^2.\end{aligned}$$

This can be written more conveniently as

$$\begin{aligned}&\sum_{x=1}^{6}(x - \overline{x})^2 f_x\\ &= \{(1-3.2)^2 \times 5\} + \{(2-3.2)^2 \times 7\} + \{(3-3.2)^2 \times 6\} + \cdots\\ &\quad + \{(6-3.2)^2 \times 3\}\\ &= 24.2 + 10.08 + 0.24 + 2.56 + 16.2 + 23.52 = 76.8.\end{aligned}$$

So the sample variance (dividing by $n - 1 = 29$) is

$$s^2 = \frac{76.8}{29} \simeq 2.65.$$

However, a theoretical probability model for the outcomes of rolls of an unbiased die is provided by the random variable X with probability mass function

$$p(x) = 1/6, \quad x = 1, 2, \ldots, 6.$$

In Example 2, you saw that the mean or expected value of X is

$$\begin{aligned}\mu = E(X) &= \left(1 \times \tfrac{1}{6}\right) + \left(2 \times \tfrac{1}{6}\right) + \left(3 \times \tfrac{1}{6}\right) + \left(4 \times \tfrac{1}{6}\right) + \left(5 \times \tfrac{1}{6}\right) + \left(6 \times \tfrac{1}{6}\right)\\ &= 3.5.\end{aligned}$$

The expected value of $(X - \mu)^2$ is found by averaging the values obtained for $(x - \mu)^2$, weighting them in the same way, according to their chance of occurrence:

$$\begin{aligned}&E\left[(X-\mu)^2\right]\\ &= \left\{(1-3.5)^2 \times \tfrac{1}{6}\right\} + \left\{(2-3.5)^2 \times \tfrac{1}{6}\right\} + \left\{(3-3.5)^2 \times \tfrac{1}{6}\right\} + \cdots\\ &\quad + \left\{(6-3.5)^2 \times \tfrac{1}{6}\right\}\\ &= \frac{6.25}{6} + \frac{2.25}{6} + \frac{0.25}{6} + \frac{0.25}{6} + \frac{2.25}{6} + \frac{6.25}{6} = \frac{17.5}{6} \simeq 2.92.\end{aligned}$$

So, comparing the population variance with the sample variance, our particular sample of 30 rolls was a little less variable than theory would have suggested.

σ^2 is pronounced 'sigma squared'. σ is the lower-case version of the Greek letter Σ, but it has nothing to do with summations!

It is a common statistical convention to denote the variance of a random variable by σ^2. The alternative notation $V(X)$ is also widely used. For discrete probability distributions, the variance is defined as follows.

The variance of a discrete distribution

For a discrete random variable X with probability mass function

$$p(x) = P(X = x)$$

and mean $\mu = E(X)$, the (**population**) **variance of X** is given by

$$\sigma^2 = V(X) = E\left[(X - \mu)^2\right] = \sum_x (x - \mu)^2 \, p(x), \tag{14}$$

where the sum is taken over all values in the range of X.

A calculation that ends in a negative variance must be wrong!

Notice that, as with the sample variance, the population variance of a discrete distribution is the sum of non-negative quantities – squared quantities times values of a p.m.f. – and so cannot be negative.

The sample standard deviation is the square root of the sample variance. Similarly, the population standard deviation is defined to be the square root of the population variance. It too is non-negative. The population standard deviation is denoted by σ, or alternatively by $S(X)$. So if X is a discrete random variable with probability mass function $p(x)$, then

$$S(X) = \sigma = \sqrt{V(X)} = \sqrt{\sum_x (x - \mu)^2 \, p(x)}.$$

So, for example, the population standard deviation of the score obtained when an unbiased die is rolled as in Example 8 is $\sqrt{17.5/6} \simeq 1.71$.

Activity 18 *The variance of the score on a biased die*

Table 10 The p.m.f. for a die with two faces showing five

w	1	3	4	5	6
$p(w)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$

In Activity 1, you found that the mean score on a die with two five-spot faces and no two-spot face is 4. Calculate the variance of W, the score obtained when such a die is rolled. (The p.m.f. is repeated alongside for your convenience.) How does this compare with the variance for an unbiased die?

The units of the variance, in both sample and population versions, and for both discrete and continuous data and distributions, are the square of the units of the original variables. So, for example, if the quantity of interest is measured in metres (m), the variance is given in metres squared (m^2). This contrasts with the standard deviation, whose units are the original units (metres in the example).

3.2 Families of distributions: the variance

The variance of a Bernoulli distribution

Recall that the possible values of a Bernoulli random variable X are 0 and 1 so, applying the definition in Equation (14), the variance of a Bernoulli distribution is given by

$$\sigma^2 = V(X) = (0-\mu)^2\, p(0) + (1-\mu)^2\, p(1).$$

Since $p(0) = 1-p$, $p(1) = p$ and, from Equation (4), the mean is p, it follows that

$$\begin{aligned}\sigma^2 &= (0-p)^2(1-p) + (1-p)^2 p \\ &= p^2(1-p) + (1-p)^2 p \\ &= p(1-p)(p+1-p) = p(1-p).\end{aligned}$$

The variance of a Bernoulli distribution

If the random variable X has a Bernoulli distribution with parameter p, $X \sim \text{Bernoulli}(p)$, then

$$\sigma^2 = V(X) = p(1-p). \tag{15}$$

Activity 19 *Variances of Bernoulli distributions*

(a) What is the variance of the score resulting from the toss of a fair coin, if we score 1 for heads and 0 for tails?

(b) Suppose that the random variable X is defined to take the value 1 when an unbiased die shows a 3 or a 6, and 0 otherwise. Find the variance of X.

You found the corresponding means in Activity 4.

The variance of a binomial distribution

To obtain the formula for the mean of a binomial distribution in Subsection 1.2, Equation (5) that relates to the mean of a sum of random variables was used. This states that the mean of a sum of random variables is equal to the sum of the means of the random variables:

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

This result holds for *any* random variables, whatever their distributions.

To obtain a formula for the variance of a binomial distribution, a general result for variances can be used. This states that the variance of a sum of *independent* random variables is equal to the sum of the variances of the individual random variables; that is, if $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n). \tag{16}$$

This result will not be proved in this module.

There was brief discussion of independence in Section 1 of Unit 2 and at the start of Section 2 of Unit 3.

Note the inclusion of the independence condition here: this is *essential*. What it means by saying that random variables are independent has not been defined formally. However, it is straightforward to describe independence of random variables informally: two or more random variables are independent if the value taken by any one of them is not influenced by the values taken by the others.

The binomial distribution is a model for the total number of successes in a sequence of independent Bernoulli trials: the outcome of each trial does not influence the outcome of any other trial, so the random variables which represent the outcomes of the Bernoulli trials are independent. Therefore, if the random variable X has a binomial distribution, $X \sim B(n, p)$, then X is the sum of n independent Bernoulli(p) random variables. Hence Equation (16) can be used to find the variance of X as follows: from Equation (15), the variance of each Bernoulli(p) random variable is $p(1-p)$, so

$$V(X) = \underbrace{p(1-p) + p(1-p) + \cdots + p(1-p)}_{n \text{ terms}} = np(1-p).$$

The variance of a binomial distribution

If the random variable X has a binomial distribution with parameters n and p, $X \sim B(n, p)$, then the variance of X is

$$\sigma^2 = V(X) = np(1-p). \tag{17}$$

Activity 20 *Variances of binomial distributions*

You found the corresponding means in Activity 5.

(a) The probability that an item from a production line is defective is 0.01. What is the variance of the number of defective items in a sample of 100 items taken from the production line?

(b) The probability that an archer hits the centre of the target with each arrow that she shoots is 3/4. Find the variance of the number of arrows that hit the centre of the target in 10 shots.

(c) The probability that a tennis player wins each match he plays against a friend is 0.7. Find the variance of his number of wins if he plays five matches.

Are these two still friends?

The variance of a geometric distribution

A formula for the variance of a geometric random variable is given without proof as follows; the details of the algebra necessary to derive the result have been omitted.

The variance of a geometric distribution

If the random variable X has a geometric distribution with parameter p, $X \sim G(p)$, then the variance of X is

$$\sigma^2 = V(X) = \frac{1-p}{p^2}. \tag{18}$$

Activity 21 *Variances of geometric distributions*

(a) Find the variance of the number of times a fair coin needs to be tossed to obtain heads.

(b) Find the variance of the number of times an unbiased six-sided die needs to be rolled to obtain a six.

The variance of a Poisson distribution

The formula for the variance of a Poisson random variable is given in the box that follows. The proof involves an extension of the argument that led to the formula for the Poisson mean in Subsection 1.2 and results from each of Subsections 5.1 and 5.2 below, so you are spared the details.

The variance of a Poisson distribution

If the random variable X has a Poisson distribution with parameter λ, $X \sim \text{Poisson}(\lambda)$, then the variance of X is

$$\sigma^2 = V(X) = \lambda. \tag{19}$$

Activity 22 *Variances of Poisson distributions*

(a) If X is the number of claims on a motor insurance policy over a 5-year period and follows the Poisson distribution with parameter $\lambda = 0.5$, what is $V(X)$?

(b) If Y is the number of yeast cells found in a randomly chosen small square on a microscope slide and follows the Poisson distribution with $\lambda = 0.6825$, what is $V(Y)$?

As you will have noticed, the Poisson distribution has the property that

$$E(X) = V(X) = \lambda;$$

that is, the mean of a Poisson distribution is equal to the variance of a Poisson distribution. The fact that the mean and variance of a Poisson

distribution are equal suggests a possible way of checking whether a Poisson model is worth considering for a particular dataset. Whenever a Poisson distribution is proposed and a sample of data is available, a quick check can be made by calculating the sample mean and sample variance. If they are close, then a Poisson model may be a good one.

Also, because its mean and variance are equal, the Poisson distribution is sometimes said to be *equi-dispersed*. Discrete distributions for which the variance is less than the mean are said to be *under-dispersed*, and those for which the variance is greater than the mean are said to be *over-dispersed*. Over-dispersion seems to be more common in practice than under-dispersion. When substantial over-dispersion is observed in a sample, it is a good indicator that a Poisson distribution is not a good model for those data, and a model allowing greater variance relative to the mean is required instead.

The dispersion effect in digital photography

Example 9 *Is the geometric distribution under-, equi- or over-dispersed?*

For a geometric random variable, $X \sim G(p)$, Equations (7) and (18) give that $E(X) = 1/p$ and $V(X) = (1-p)/p^2$, respectively. Notice that

$$V(X) = \frac{1-p}{p} \times \frac{1}{p} = \frac{1-p}{p} \times E(X).$$

So in this case we will have:

- $V(X) < E(X)$ and hence under-dispersion when $(1-p)/p < 1$
- $V(X) = E(X)$ and hence equi-dispersion when $(1-p)/p = 1$
- $V(X) > E(X)$ and hence over-dispersion when $(1-p)/p > 1$.

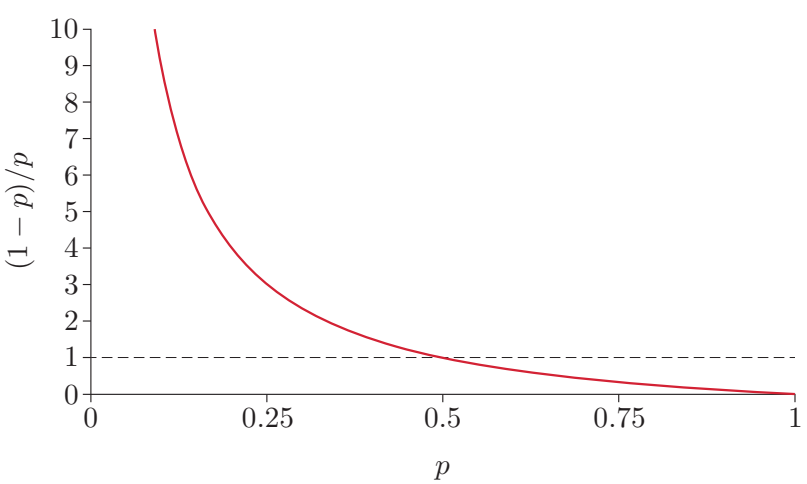


Figure 5 The function $(1-p)/p$; the dashed line is at height 1

A graph of the function $f(p) = (1-p)/p$ is given in Figure 5, from which we can read off the under-/equi-/over-dispersion situation for the geometric distribution in terms of its parameter p:

- $(1-p)/p < 1$ when $p > 1/2$, so the geometric distribution is under-dispersed when $p > 1/2$

- $(1-p)/p = 1$ when $p = 1/2$, so the geometric distribution is equi-dispersed when $p = 1/2$
- $(1-p)/p > 1$ when $p < 1/2$, so the geometric distribution is over-dispersed when $p < 1/2$.

The situation is simpler both to derive and to state for the binomial distribution. You can do this for yourself in the next activity.

Activity 23 *Is the binomial distribution under-, equi- or over-dispersed?*

Let $X \sim B(n, p)$ where $0 < p < 1$. For such a random variable, Equations (6) and (17) give that $E(X) = np$ and $V(X) = np(1-p)$, respectively. For what values of n and p is the binomial distribution under-dispersed, equi-dispersed and over-dispersed?

The variance of a discrete uniform distribution

The variance of a discrete uniform distribution may be obtained by algebraic manipulation but the details of its derivation will not be included here. The result is given in the following box.

The variance of a discrete uniform distribution

If the random variable X has a discrete uniform distribution on $m, m+1, \ldots, n$, then the variance of X is

$$V(X) = \frac{1}{12}(n-m)(n-m+2). \tag{20}$$

Activity 24 *Variances of discrete uniform distributions*

(a) Use Equation (20) to find the variance of the score obtained when an unbiased six-sided die is rolled.

(b) If Y is the random variable representing the number chosen when an integer between 1 and 9 is selected at random, use Equation (20) to find the variance of Y.

(c) If Z is the random variable representing the number chosen when an integer between 0 and 9 is selected at random, what is the variance of Z? How does $V(Z)$ compare with $V(Y)$ obtained in part (b)?

Bowling scores: probably not uniform on $0, 1, \ldots, 10$?

Exercises on Section 3

Exercise 7 *Practice at finding variances*

Table 11 The p.m.f. of X

x	2	3	4	5
$p(x)$	0.1	0.2	0.3	0.4

Table 12 The p.m.f. of Y

y	0	1	2	3	4
$p(y)$	0.4	0.2	0.1	0.1	0.2

(a) The probability distribution of a random variable X was given in Table 7, repeated as Table 11 for convenience. You showed in Exercise 1(a) that the mean of X is 4. Find the variance of X.

(b) The probability distribution of a random variable Y was given in Table 8, repeated as Table 12 for convenience. You showed in Exercise 1(b) that the mean of Y is 1.5. Find the variance of Y.

Exercise 8 *Finding variances for standard distributions*

In Exercise 3 you found the means of three random variables associated with an archer whose probability of hitting the centre of the target with each arrow she shoots is 0.8.

(a) Find the variance of X, the random variable which takes the value 1 when the archer hits the centre of the target and 0 when she misses.

(b) If the archer shoots seven arrows, find the variance of the number of arrows that hit the centre of the target.

(c) If the archer continues to shoot arrows at the centre of the target until she misses, find the variance of the number of arrows she shoots.

4 The variance of a continuous distribution

In Section 2, the formula defining the mean of a continuous random variable was obtained from that for a discrete random variable by replacing the p.m.f. by a p.d.f., and the sum by an integral. The formula defining the variance of a continuous random variable may be obtained using the same approach.

The variance of a continuous random variable

For a continuous random variable X with probability density function $f(x)$ and mean $\mu = E(X)$, the **(population) variance of X** is given by

$$\sigma^2 = V(X) = E\left[(X-\mu)^2\right] = \int (x-\mu)^2 f(x)\,dx, \qquad (21)$$

where the integral is taken over all values x in the range of X.

Again, the population variance σ^2 cannot be negative. Also, the population standard deviation σ is the square root of the population variance, $\sigma = \sqrt{V(X)}$, and it cannot be negative either.

Example 10 *The variance of a continuous uniform distribution*

Let us calculate the variance of a random variable, X, following the continuous uniform distribution with probability density function

$$f(x) = \frac{1}{b-a}, \quad a < x < b.$$

You saw in Section 2 that the mean associated with this distribution is $\mu = E(X) = (a+b)/2$. The variance is therefore given by

You need not worry about the details of the tedious algebraic manipulations.

$$\begin{aligned}
V(X) = E\left[(X-\mu)^2\right] &= \int_a^b (x-\mu)^2 f(x)\,dx \\
&= \int_a^b \left\{x - \frac{a+b}{2}\right\}^2 \frac{1}{b-a}\,dx \\
&= \frac{1}{b-a}\int_a^b \left\{x^2 - (a+b)x + \frac{(a+b)^2}{4}\right\} dx \\
&= \frac{1}{b-a}\left[\frac{x^3}{3} - \frac{(a+b)x^2}{2} + \frac{(a+b)^2 x}{4}\right]_a^b \\
&= \frac{1}{b-a}\left\{\frac{b^3}{3} - \frac{(a+b)b^2}{2} + \frac{(a+b)^2 b}{4} - \left(\frac{a^3}{3} - \frac{(a+b)a^2}{2} + \frac{(a+b)^2 a}{4}\right)\right\} \\
&= \frac{b^3 - 3b^2a + 3a^2b - a^3}{12(b-a)} = \frac{(b-a)^3}{12(b-a)} \\
&= \frac{(b-a)^2}{12}.
\end{aligned}$$

This result completes your collection of highlighted means and variances of families of distributions in this unit.

The variance of a continuous uniform distribution

If the random variable X has a continuous uniform distribution on (a, b), then the variance of X is given by

$$V(X) = \frac{1}{12}(b-a)^2. \tag{22}$$

Activity 25 *Variances and standard deviations of continuous uniform distributions*

Use the result just obtained to answer the following questions.

(a) In Activity 26 of Unit 3, it was argued that the position, X, of a fault in a cable of length 40 metres might be modelled by a continuous uniform distribution on $(0, 40)$. In Activity 17(a) you showed that the mean position of a fault was at 20 metres. What is the variance of the position of the fault?

(b) What is the variance of a random variable Y which follows the standard continuous uniform distribution with $a = 0$, $b = 1$? What is the standard deviation of Y?

(c) Someone claims that the standard deviation of the $U(a, b)$ distribution is $(a - b)/\sqrt{12}$. Explain why this cannot be correct, and give the correct formula for the standard deviation.

You can evaluate variances for other continuous distributions, too.

Activity 26 *The variance of another distribution*

In Activity 14(a), you showed that the mean of the random variable X following the distribution with p.d.f.

$$f(x) = 3x^2, \quad 0 < x < 1,$$

is $E(X) = 3/4$. Find the variance of this distribution.

Example 11 *The variance of journey time?*

In Example 5, the following model for a man's journey to work was considered:

$$f(x) = \frac{1}{5} - \frac{x}{250}, \quad 20 < x < 30.$$

The mean journey time, μ, was calculated to be 74/3 minutes (or 24 minutes and 40 seconds). What about the variance of the man's journey times? Well,

$$\begin{aligned} V(X) &= \int_{20}^{30} (x-\mu)^2 f(x)\,dx = \int_{20}^{30} \left(x - \frac{74}{3}\right)^2 \left(\frac{1}{5} - \frac{x}{250}\right) dx \\ &= \int_{20}^{30} \left(x^2 - \frac{148x}{3} + \frac{5476}{9}\right)\left(\frac{1}{5} - \frac{x}{250}\right) dx \\ &= \int_{20}^{30} \left(\frac{x^2}{5} - \frac{148x}{15} + \frac{5476}{45} - \frac{x^3}{250} + \frac{74x^2}{375} - \frac{2738x}{1125}\right) dx \\ &= \int_{20}^{30} \left(\frac{5476}{45} - \frac{13\,838x}{1125} + \frac{149x^2}{375} - \frac{x^3}{250}\right) dx = \cdots \end{aligned}$$

Hold on, hold on! The algebra is getting out of hand even for what is quite a simple distribution. (It wasn't exactly easy even for the continuous uniform distribution in Example 10!) Surely there's a better way. Well, happily, there is. The calculation of variances is usually done in a different manner that will be introduced in Subsection 5.2, where, among other things, the calculation of the variance for the journey time model will be completed.

Exercise on Section 4

Exercise 9 *The variance of another simple distribution*

In Exercise 4(a), you showed that the mean of the random variable X following the distribution with p.d.f.

$$f(x) = 4x^3, \quad 0 < x < 1,$$

is $E(X) = 4/5$. Find the variance of this distribution.

5 Means and variances of linear functions

5.1 Linear functions of a random variable

The three examples below illustrate the need for some further results on means and variances of random variables. They introduce random variables that are in various ways simply related to random variables whose means and variances are already known. What are the effects of these relationships on the mean and variance of the related random variable?

Example 12 *Chest measurements*

Let X be a random variable representing the chest measurement of a British male as made sometime in the early part of the twentieth century. Such measurements were recorded in inches. Suppose that a reasonable model for such chest measurements has (population) mean $\mu = 40$ inches and (population) standard deviation $\sigma = 2$ inches. Suppose that it is desired to make a comparison with present-day British males, whose chest measurements are recorded in centimetres. Since 1 inch is approximately equal to 2.54 centimetres, it is straightforward to convert an individual measurement from inches to centimetres: if Y is the random variable representing the chest measurement of a British male in the early part of the twentieth century in centimetres, then

$$Y = 2.54X.$$

But what can we say about the distribution of chest measurements in centimetres if we know about the distribution of chest measurements in inches? In particular, what are the mean and standard deviation of Y, and how are they related to the mean and standard deviation of X?

Example 13 *Geometric distribution starting from zero*

In Subsection 3.1 of Unit 3, the geometric distribution was introduced. It is the distribution of a random variable X representing the number of trials from the start of a sequence of independent Bernoulli trials up to *and including* the first success. So

$X = 1$ if the first trial is a success;

$X = 2$ if the first trial is a failure and the second is a success;

$X = 3$ if the first two trials are failures and the third is a success;

and so on. In some situations, it is more appropriate to consider a different random variable, Y say, representing the number of trials from the start of a sequence of independent Bernoulli trials up to *but not including* the first success. So

$Y = 0$ if the first trial is a success;

$Y = 1$ if the first trial is a failure and the second is a success;

$Y = 2$ if the first two trials are failures and the third is a success;

and so on. Or, more succinctly,

$$Y = X - 1.$$

It is certainly possible to derive the distribution of Y directly and then to work out its mean and variance using that distribution. But we know, from Sections 1 and 3, that $E(X) = 1/p$ and $V(X) = (1-p)/p^2$, where p is the probability of success on a single Bernoulli trial. Presumably, we should be able to obtain the mean and variance of Y from these without going to the effort of first deriving the distribution of Y?

Example 14 *Temperature scales*

In the UK, weather forecasters give temperatures in degrees Celsius (°C), a scale in which 0 ° is the freezing point of water and 100 ° is the boiling point of water. In the USA, weather forecasters give temperatures in degrees Fahrenheit (°F), a scale in which 32 ° is the freezing point of water and 212 ° is the boiling point of water. Let X be a random variable representing a temperature given in °C, and let Y be a random variable representing the same temperature given in °F. The two temperature scales are related by

$$Y = \frac{9}{5}X + 32. \tag{23}$$

So if we know the mean temperature on the Celsius scale, what is the mean temperature on the Fahrenheit scale? If we know the variance of the temperature on the Celsius scale, what is the variance of the temperature on the Fahrenheit scale?

Let X denote any of the 'original' random variables in Examples 12 to 14, and let Y denote the corresponding related variable. Then the common

structure is that, in each case, $Y = aX + b$ where a and b are constants. That is, Y is a linear function of X.

Activity 27 *Values of a and b*

Confirm the claim just made by identifying the values of a and b in the representation $Y = aX + b$ for each of Examples 12 to 14.

So the problem is: if $Y = aX + b$ and we know $E(X)$ and $V(X)$, what are $E(Y)$ and $V(Y)$?

You may well be able to guess the answer in the case that $a = 1$ so that Y is X shifted by the value b; you are asked to think about this in the next activity.

Activity 28 *The effect of adding b*

Consider $Y = X + b$.

(a) If $E(X) = \mu$, what do you think $E(Y)$ is?

(b) If $V(X) = \sigma^2$, what do you think $V(Y)$ is?

In order to avoid confusion, it is stressed that in this section we are concerned with the mean and variance of a linear function of a *single* random variable, X say. This is a different question than the mean and variance of a linear function of *two or more* random variables, such as, for example, the quantities $E(X_1 + X_2 + \cdots + X_n)$ and $V(X_1 + X_2 + \cdots + X_n)$, formulas for which were used earlier in this unit.

The mean of $Y = aX + b$

The derivation of the mean of a linear function $Y = aX + b$ of a random variable X is quite straightforward. Details are included for the discrete case.

In Section 1, the mean or expected value of a discrete random variable X was defined as

$$E(X) = \sum_x x\, p(x),$$

where $p(x)$ is the probability mass function and the sum is taken over all values in the range of X.

More generally, the mean or expected value of a function $h(X)$ of a discrete random variable X is defined by

$$E[h(X)] = \sum_x h(x)\, p(x).$$

For example, $h(x) = (x - \mu)^2$ in the formula for the variance in Equation (14).

This definition makes sense as an average of the different values that $h(X)$ may take, weighted according to their chance of occurrence: $h(X)$ takes the value $h(x)$ whenever X takes the value x.

Now suppose that $Y = aX + b$, where X is a discrete random variable and a and b are constants. Using the definition above gives

$$E(Y) = E(aX + b) = \sum_x (ax + b)\, p(x).$$

The desired result follows on splitting this sum into two sums:

$$\begin{aligned} E(aX + b) &= \sum_x ax\, p(x) + \sum_x b\, p(x) \\ &= a \sum_x x\, p(x) + b \sum_x p(x) \\ &= a\, E(X) + b. \end{aligned}$$

The last line follows from the definition of $E(X)$ and because the sum of the probabilities $p(x)$ is equal to 1 for any discrete probability distribution.

The result also holds when X is a continuous random variable. This can be established using similar arguments, with integrals instead of sums but, for brevity, the derivation in the continuous case will not be included. Hence we have the following result which holds for any random variable X, whether discrete or continuous.

The mean of a linear function of a random variable

If X is a random variable and a and b are constants, then the mean of the random variable $Y = aX + b$ is given by

$$E(Y) = E(aX + b) = a\, E(X) + b. \tag{24}$$

Notice that when $a = 1$,

$$E(X + b) = E(X) + b,$$

as argued in Activity 28(a). Similarly, if $b = 0$, then

$$E(aX) = a\, E(X),$$

which is reasonable by a similar argument: if a random variable X is multiplied by a constant a, then all values of X are multiplied by a, so you would expect the mean value of X also to be multiplied by a.

Example 15 *The mean chest measurement in centimetres*

In Example 12, a distribution with mean 40 was suggested as a model for the random variable X which represents the chest measurement of an early twentieth-century male in inches. What is the mean of Y, the corresponding random variable which represents the chest measurement of an early twentieth-century male in centimetres?

Using Equation (24) with $a = 2.54$ and $b = 0$, the random variable $Y = 2.54X$ has mean

$$E(Y) = E(2.54X) = 2.54E(X) = 2.54 \times 40 = 101.6 \text{ cm}.$$

Activity 29 *The mean of a geometric distribution starting from zero*

In Example 13, you were reminded that if X follows a geometric distribution with parameter p, then $E(X) = 1/p$. The random variable $Y = X - 1$ was then introduced; what is $E(Y)$?

Activity 30 *The mean water temperature*

Larvae of *Culex* mosquito in water

According to the World Health Organization, in 2015 there were 214 million malaria cases reported worldwide (although this figure is 37% less than the one in 2000). One of the factors in the spread of malaria is the density of the spreading agent – the *Anopheles* mosquito – in the population. An environmental feature that in turn affects this density is the water temperature where the mosquito larvae develop. This will be subject to variation. Suppose that the water temperature measured in °C is represented by a random variable X. If Y is a random variable representing the water temperature measured in °F, then Equation (23) gives

$$Y = \frac{9}{5}X + 32.$$

Suppose that the probability distribution of the water temperature in degrees Celsius has mean 26. Find the mean water temperature in degrees Fahrenheit.

The variance of $Y = aX + b$

The variance of any random variable Y is defined by

$$V(Y) = E\big[(Y - \mu)^2\big] = E\big[\{Y - E(Y)\}^2\big].$$

If Y is a function $h(X)$ of X, this formula still applies:

$$V\{h(X)\} = E\big([h(X) - E\{h(X)\}]^2\big).$$

In particular, if $Y = aX + b$, then

$$V(Y) = V(aX + b) = E\big[\{aX + b - E(aX + b)\}^2\big].$$

But, by Equation (24), $E(aX + b) = a\,E(X) + b$. So

$$\begin{aligned} V(Y) &= E\big[\{aX + b - a\,E(X) - b\}^2\big] \\ &= E\big[\{aX - a\,E(X)\}^2\big] = E\big[a^2\{X - E(X)\}^2\big]. \end{aligned}$$

In this expression $\{X - E(X)\}^2$ is a random variable, W say, and a^2 is a constant. So, applying Equation (24) again, we have $E(a^2W) = a^2E(W)$. That is, a^2 can be taken outside the brackets above. Hence we have

$$V(Y) = a^2E\big[\{X - E(X)\}^2\big] = a^2\,V(X).$$

This argument is simpler than perhaps it looks. A version that tries to de-clutter the argument by de-emphasising the brackets is given in the following screencast.

Screencast 4.4 Obtaining the formula for the variance of a linear function of X

The above argument is valid for both discrete and continuous random variables, so we have the following general result for the variance of a linear function of a random variable.

The variance of a linear function of a random variable

If X is a random variable and a and b are constants, then the variance of the random variable $Y = aX + b$ is given by

$$V(Y) = V(aX + b) = a^2\,V(X). \tag{25}$$

Key here is being able to apply the result, rather than being able to prove it.

Notice that the constant b does not feature in the right-hand side of Equation (25). This confirms the discussion in Activity 28, that adding a constant b to a random variable adds b to the mean but does not change the variability about the mean.

A formula for the standard deviation of $aX + b$ can be found by taking the square root of each side of Equation (25):

$$S(aX + b) = \sqrt{a^2\,V(X)} = |a|\,S(X). \tag{26}$$

The modulus or absolute value symbol in Equation (26) is important: the constant a could be negative, but a standard deviation is always non-negative. (If a random variable X is multiplied by a constant a, then values of X are spread out by a factor $|a|$, that is, by a factor of a if $a > 0$, and by a factor of $-a$ if $a < 0$.)

Example 16 *The variance and standard deviation of chest measurements in centimetres*

In Example 12, a distribution with mean 40 and standard deviation 2 was suggested as a model for the random variable X which represents the chest measurement of an early twentieth-century male in inches. What is the variance of Y, the corresponding random variable which represents the chest measurement of an early twentieth-century male in centimetres? What is the standard deviation of Y?

The variance of X is $2^2 = 4$ square inches and $Y = 2.54X$ cm. So, using Equation (25),

$$V(Y) = 2.54^2\,V(X) = 2.54^2 \times 4 \simeq 25.81 \text{ cm}^2.$$

Using Equation (26), the standard deviation of Y is

$$S(Y) = 2.54\,S(X) = 2.54 \times 2 = 5.08 \text{ cm}.$$

Activity 31 *The variance of a geometric distribution starting from zero*

In Example 13, you were reminded that if X follows a geometric distribution with parameter p, then $V(X) = (1-p)/p^2$. The random variable $Y = X - 1$ was then introduced; what is $V(Y)$?

Activity 32 *The variance and standard deviation of water temperature*

Activity 30 concerned water temperature in areas in which mosquitos that spread malaria are endemic. Suppose now that the probability distribution of the water temperature in degrees Celsius, X, has mean 26 and standard deviation 1.25. Find the standard deviation and the variance of the water temperature in degrees Fahrenheit, Y, using the relationship $Y = \frac{9}{5}X + 32$.

5.2 An alternative formula for the variance of a random variable

The variance $V(X)$ of a random variable X with mean $E(X)$ was defined for discrete distributions in Section 3 and for continuous distributions in Section 4 by

$$V(X) = E\big[(X-\mu)^2\big] = E\big[(X - E(X))^2\big]. \tag{27}$$

It is the way that the expected value is calculated that differs between discrete and continuous cases.

This formula is not always simple to apply. This was stressed in Example 11 where, even for the relatively simple model $f(x) = \frac{1}{5} - \frac{x}{250}$, $20 < x < 30$, the algebraic manipulations seemed to be getting out of hand and were abandoned. In addition, if a rounded value is used for the mean to simplify the calculations, then serious rounding error may be introduced in calculating $E\big[(X-\mu)^2\big]$ directly.

A useful alternative formula for the variance may be derived using Equation (24) for the mean of a linear function of a random variable. As you will see, this formula is often simpler to use in practice than Equation (27) directly. To obtain this formula, we need the incidental result that if $h_1(X)$ and $h_2(X)$ are two functions of a random variable X, then the expected value of their sum is equal to the sum of their expected values:

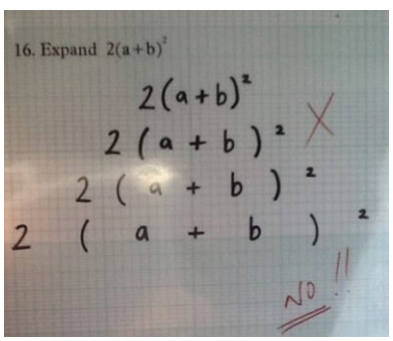
$$E[h_1(X) + h_2(X)] = E[h_1(X)] + E[h_2(X)]. \tag{28}$$

This is another manifestation of the term 'linearity of expectation'.

It is straightforward to obtain this result. For example, if X is discrete with probability mass function $p(x)$, then

$$\begin{aligned} E[h_1(X) + h_2(X)] &= \sum_x (h_1(x) + h_2(x))\, p(x) \\ &= \sum_x \{h_1(x)\, p(x) + h_2(x)\, p(x)\} \\ &= \sum_x h_1(x)\, p(x) + \sum_x h_2(x)\, p(x) \\ &= E[h_1(X)] + E[h_2(X)]. \end{aligned}$$

A similar argument, but with integrals instead of sums, confirms the result when X is a continuous random variable.



Now, expanding the square on the right-hand side of the formula for the variance in Equation (27) gives

$$V(X) = E(X^2 - 2\mu X + \mu^2).$$

The expression $X^2 - 2\mu X + \mu^2$ may be written in the form $h_1(X) + h_2(X)$ where $h_1(X) = X^2$ and $h_2(X) = -2\mu X + \mu^2$. Then using Equation (28) gives

$$V(X) = E(X^2) + E(-2\mu X + \mu^2).$$

Furthermore, applying Equation (24) for the mean of a linear function of X to the second term on the right gives

$$E(-2\mu X + \mu^2) = -2\mu\, E(X) + \mu^2.$$

Since $E(X) = \mu$, this gives

$$E(-2\mu X + \mu^2) = -2\mu^2 + \mu^2 = -\mu^2$$

and hence

$$V(X) = E(X^2) - \mu^2.$$

Note that this argument holds for both discrete and continuous random variables, so the formula may be used for any random variable.

This formula for the variance of a random variable X is very useful as it is often easier to apply than that in Equation (27) which you have used up to now: calculating $E(X^2)$ is usually easier than calculating $E\left[(X-\mu)^2\right]$. In fact, $E(X^2)$ is a special case of $E[h(X)]$ mentioned earlier and so is given by

$$E(X^2) = \sum_x x^2\, p(x)$$

in the discrete case and

$$E(X^2) = \int x^2 f(x)\, dx$$

in the continuous case. The result is stated formally in the box below.

An alternative formula for the variance of a random variable

If X is a random variable with mean $\mu = E(X)$, then the variance of X may be obtained using the formula

$$V(X) = E(X^2) - \mu^2. \tag{29}$$

That is,

$$V(X) = E\left[(X-\mu)^2\right] = E(X^2) - \mu^2.$$

You might remember this as

[the variance] equals [the mean of the squares] minus [the square of the mean].

Example 17 *The variance of the score on an unbiased die again*

Suppose that X is a random variable representing the score obtained when an unbiased six-sided die is rolled. Then X has p.m.f.

$$p(x) = 1/6, \quad x = 1, 2, \ldots, 6.$$

In Example 2, you saw that the mean score is $E(X) = \mu = 3.5$, while in Example 8, Equation (14) was used to show that the variance of X is approximately 2.92. Using the alternative Equation (29) to find the variance involves first finding the value of $E(X^2)$ and then subtracting the square of the mean. The expected value of X^2 is

$$\begin{aligned} E(X^2) &= \left(1^2 \times \tfrac{1}{6}\right) + \left(2^2 \times \tfrac{1}{6}\right) + \cdots + \left(6^2 \times \tfrac{1}{6}\right) \\ &= \frac{1+4+9+16+25+36}{6} = \frac{91}{6}. \end{aligned}$$

So

$$V(X) = E(X^2) - \mu^2 = \frac{91}{6} - 3.5^2 \simeq 2.92,$$

as obtained previously using Equation (14).

Activity 33 *Using the alternative formula in another discrete case*

The probability distribution of the random variable Y is given in Table 13.

Table 13 The p.m.f. of Y

y	0	1	2	3
$p(y)$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{3}{7}$	$\frac{1}{7}$

(a) Find the mean of Y.

(b) Use Equation (29) to find the variance of Y.

Equation (29) is also very effective in simplifying the calculations for obtaining the variance in Example 10 (for the continuous uniform distribution) and Example 11 (for a model for journey times). You can rework Example 10 for yourself in the next activity, but first let's take another look at the variance of the model for journey times.

Example 18 *The variance of journey time revisited*

In Examples 5 and 11, the following model for a man's journey to work was considered:

$$f(x) = \frac{1}{5} - \frac{x}{250}, \quad 20 < x < 30.$$

The mean journey time according to the model, μ, was calculated to be $74/3$ minutes in Example 5. But calculation of the variance of the man's journey times was abandoned in Example 11 when the algebra became hard. Let's try again, this time calculating $E(X^2)$ – and then subtracting μ^2 – instead of calculating $E\left[(X-\mu)^2\right]$ directly. We have

$$\begin{aligned} E(X^2) &= \int_{20}^{30} x^2 f(x)\,dx = \int_{20}^{30} x^2\left(\frac{1}{5} - \frac{x}{250}\right)dx \\ &= \int_{20}^{30}\left(\frac{x^2}{5} - \frac{x^3}{250}\right)dx = \left[\frac{x^3}{15} - \frac{x^4}{1000}\right]_{20}^{30} \\ &= \frac{27\,000}{15} - \frac{810\,000}{1000} - \left(\frac{8000}{15} - \frac{160\,000}{1000}\right) \\ &= 1800 - 810 - \left(\frac{1600}{3} - 160\right) = 1150 - \frac{1600}{3} = \frac{1850}{3}. \end{aligned}$$

Therefore

$$V(X) = E(X^2) - \mu^2 = \frac{1850}{3} - \left(\frac{74}{3}\right)^2 = \frac{74}{9} \simeq 8.22.$$

So now he can go!

Activity 34 *The variance of the continuous uniform distribution revisited*

The continuous uniform distribution has probability density function

$$f(x) = \frac{1}{b-a}, \quad a < x < b,$$

and, as shown in Section 2, $\mu = \frac{1}{2}(a+b)$. Use Equation (29) to confirm that $\sigma^2 = (b-a)^2/12$. Hint: you might find the algebraic equality $(b^3 - a^3)/(b-a) = a^2 + ab + b^2$ (which holds for $a < b$) useful.

Activity 35 *The variance and standard deviation of the lengths of brown trout fry*

In Activity 15, the distribution with probability density function

$$f(x) = \frac{1}{30}(10x - x^2 - 14), \quad 3 < x < 6,$$

was used for the lengths (in cm) of brown trout fry in a hatchery pond. The mean length of the brown trout fry according to the model was shown to be 4.575 cm. Find the variance of the lengths of these brown trout fry according to the model, and hence find the standard deviation.

The following screencast works through another example of calculating the variance of a continuous distribution. This screencast uses the same distribution as was used in Screencast 4.3, for which the mean was calculated.

Screencast 4.5 Obtaining the variance and standard deviation of a continuous distribution

Exercises on Section 5

Exercise 10 *Relabelled Bernoulli trials*

At the beginning of Subsection 1.1 of Unit 3, it was said that 'Where an experiment involves a Bernoulli trial, it is usual to match the number 1 to one outcome and the number 0 to the other'. Let X denote such an outcome, and let p be the probability of obtaining a '1'. Then we know from Equations (4) and (15) that $E(X) = p$ and $V(X) = p(1-p)$. The numerical labelling is, however, arbitrary. Suppose that the first outcome is still labelled '1' but the second outcome is labelled '-1'. Let Y denote the relabelled outcome, and let p remain the probability of obtaining a '1'.

(a) Give a formula for Y in terms of X.

(b) What is $E(Y)$?

(c) What is $V(Y)$?

Exercise 11 *Sections of a chemical reactor*

Variation in the section temperature in °F across the 1250 sections of a chemical reactor, T, may be assumed to be adequately modelled by a distribution with mean 845 and standard deviation 40. Let W be the random variable representing the section temperature in °C.

These are fictional data based on a real investigation. See Cox, D.R. and Snell, E.J. (1981) *Applied Statistics*, London, Chapman and Hall, p. 8.

(a) Using Equation (23), write W in terms of T.

(b) What is $E(W)$?

(c) What is $S(W)$?

Exercise 12 *The variance of a triangular distribution*

In Exercise 5, you considered a random variable X following a family of triangular distributions, indexed by parameter $b > 0$, with p.d.f.

$$f(x) = \frac{2(b-x)}{b^2}, \quad 0 < x < b,$$

and showed that $E(X) = b/3$. Find the variance of this triangular distribution (in terms of b).

Exercise 13 *The standard deviation of bulldozer return times*

In Exercise 6, you considered a model for the return times, X, in minutes, of a bulldozer when carrying out a particular earthmoving task. The model has p.d.f.

$$f(x) = \frac{15}{16\sqrt{2}}\sqrt{x}(2-x), \quad 0 < x < 2,$$

and you showed that $E(X) = 6/7$ minutes. What is the standard deviation of the bulldozer return times according to this model?

Summary

This unit has focused on means and variances of probability models for both discrete and continuous data, and some basic properties thereof. These are population analogues of the sample mean and sample variance.

The population mean, $E(X) = \mu$, is defined by

$$\mu = \sum_x x\,p(x) \quad \text{for discrete data}$$

and by

$$\mu = \int x\,f(x)\,dx \quad \text{for continuous data.}$$

The population variance, $V(X) = \sigma^2$, is defined in either case as

$$V(X) = E\big[(X-\mu)^2\big].$$

You have seen that this has an alternative formula,

$$V(X) = E(X^2) - \mu^2,$$

which is often easier to use. The population standard deviation, $S(X) = \sigma$, is defined as

$$S(X) = \sqrt{V(X)}.$$

If Y is a linear function of X, so that $Y = aX + b$, say, where a and b are constants, then you have seen that

$$E(Y) = a\,E(X) + b \quad \text{and} \quad V(Y) = a^2\,V(X).$$

In addition to a number of more specific models, formulas have been given for the mean and variance of a member of each of the six main families of distributions so far discussed: they are listed below for easy reference. These formulas are well-known results that may be quoted whenever the mean or variance of a member of one of these families is required.

Table 14 Means and variances of standard distributions

Distribution	Mean	Variance
Bernoulli (p)	p	$p(1-p)$
Binomial, $B(n,p)$	np	$np(1-p)$
Geometric, $G(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	λ	λ
Discrete uniform on $m, m+1, \ldots, n$	$\frac{1}{2}(n+m)$	$\frac{1}{12}(n-m)(n-m+2)$
Continuous uniform, $U(a,b)$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$

Learning outcomes

After you have worked through this unit, you should be able to:

- appreciate that the following terms are all equivalent: population mean, mean of a probability distribution, mean of a random variable, expected value of a random variable, expectation of a random variable
- calculate the mean, $E(X) = \mu$, of both discrete and continuous probability distributions in simple cases
- calculate the variance, $V(X) = \sigma^2 = E\left[(X-\mu)^2\right] = E(X^2) - \mu^2$, and standard deviation, $S(X) = \sqrt{V(X)} = \sigma$, of both discrete and continuous probability distributions in simple cases, appreciating that neither quantity can be negative
- appreciate that the mean of a sum of random variables is equal to the sum of their means, and that the variance of a sum of *independent* random variables is equal to the sum of their variances
- use standard results to find the mean and variance of a member of one of the following families of distributions: Bernoulli, binomial, geometric, Poisson and uniform (discrete and continuous)
- calculate the mean and variance of a linear function of a random variable
- acknowledge that the parameter p of a geometric distribution may be estimated from data by the reciprocal of the sample mean.

Solutions to activities

Solution to Activity 1

The mean score obtained when the die described in the question is rolled is

$$\mu = \left(1 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{6}\right) + \left(4 \times \frac{1}{6}\right) + \left(5 \times \frac{1}{3}\right) + \left(6 \times \frac{1}{6}\right)$$
$$= \frac{1+3+4+10+6}{6} = \frac{24}{6} = 4.$$

The effect of replacing the two-spot face on an unbiased die by a second five-spot face is to increase the mean score from 3.5 to 4.

Solution to Activity 2

(a) The total number of families in the UK with at least one dependent child in 2012 was

$$3\,700\,000 + 3\,000\,000 + 1\,100\,000 = 7\,800\,000.$$

The mean family size is therefore

$$\mu = \frac{(1 \times 3\,700\,000) + (2 \times 3\,000\,000) + (3 \times 1\,100\,000)}{7\,800\,000}$$
$$= \frac{3\,700\,000 + 6\,000\,000 + 3\,300\,000}{7\,800\,000} = \frac{13\,000\,000}{7\,800\,000} \simeq 1.67.$$

Two points to note: no family actually has the mean number of children, which is 1.67; and the calculation could have been made simpler by noticing that the 'millions' cancel from numerator and denominator, so you could have calculated the mean by

$$\mu = \frac{3.7 + 6 + 3.3}{3.7 + 3 + 1.1} \simeq 1.67.$$

(b) The effect of approximating 'three or more children' by 'three children' is to reduce the mean family size compared with what it actually is.

As it happens, the effect in this case is not large: a more exact calculation by the Office for National Statistics gives the mean family size as 1.7, which is the same as you found, correct to one decimal place.

Solution to Activity 3

The population mean of any Bernoulli distribution is

$$\mu = \sum_{x=0}^{1} x\,p(x) = 0 \times p(0) + 1 \times p(1) = 0 \times (1-p) + 1 \times p = p.$$

That is, the population mean of a Bernoulli distribution is the parameter p used to index the family of Bernoulli distributions.

Solution to Activity 4

(a) For a fair coin, $P(\text{heads}) = \frac{1}{2}$, so $p(1) = \frac{1}{2}$. Hence $p = \frac{1}{2}$ and, using Equation (4), $\mu = E(X) = \frac{1}{2}$.

(b) For this die-rolling situation, $p(1) = \frac{1}{3} = p$ and hence $\mu = E(X) = \frac{1}{3}$.

Solution to Activity 5

(a) The number of defective items in a sample of size 100 has a binomial distribution, $B(100, 0.01)$. So the mean number of defective items in a sample of size 100 is

$$np = 100 \times 0.01 = 1.$$

(b) The number of arrows that hit the centre of the target has a binomial distribution, $B(10, 3/4)$. So the mean number of arrows out of 10 that hit the centre of the target is

$$np = 10 \times \tfrac{3}{4} = 7.5.$$

(c) The number of wins in five matches has a binomial distribution, $B(5, 0.7)$. So the player's expected number of wins in five matches is

$$np = 5 \times 0.7 = 3.5.$$

Solution to Activity 7

Equation (7) gives the following means.

(a) $1/\frac{1}{2} = 2$. (b) $1/\frac{1}{3} = 3$. (c) $1/\frac{1}{6} = 6$.

Were these the values you came up with in Activity 6?

Solution to Activity 8

(a) The proportion of darts that hit the bull's-eye is $12/50 = 0.24$. So an estimate of p is 0.24.

(b) The total number of darts that hit the bull's-eye out of the 200 throws is $4 + 10 + \cdots + 4 = 54$, so an estimate of p is $54/200 = 0.27$.

(c) The sample mean is

$$\frac{(1 \times 8) + (2 \times 5) + \cdots + (15 \times 1)}{30} = \frac{120}{30} = 4.$$

So an estimate of p is $1/4 = 0.25$.

Solution to Activity 9

Equation (8) gives the following means.

(a) $E(X) = 0.5$. (b) $E(Y) = 0.6825$.

Solution to Activity 10

(a) If X is the score obtained when an unbiased tetrahedral die is rolled, then the mean score is

$$E(X) = \left(1 \times \frac{1}{4}\right) + \left(2 \times \frac{1}{4}\right) + \left(3 \times \frac{1}{4}\right) + \left(4 \times \frac{1}{4}\right)$$
$$= \frac{1+2+3+4}{4} = \frac{10}{4} = 2.5.$$

(b) If Y is the digit chosen, then

$$E(Y) = \left(1 \times \frac{1}{9}\right) + \left(2 \times \frac{1}{9}\right) + \cdots + \left(9 \times \frac{1}{9}\right)$$
$$= \frac{1+2+\cdots+9}{9} = \frac{45}{9} = 5.$$

(c) It looks as though the mean is midway between the lowest and highest values. A formula is derived in the next activity.

Solution to Activity 11

The mean of a random variable X which has a discrete uniform distribution with parameter n is given by

$$E(X) = \sum_{x=1}^{n} x\,p(x) = \sum_{x=1}^{n} x\left(\frac{1}{n}\right) = \frac{1}{n}\sum_{x=1}^{n} x,$$

which, from Equation (10), becomes

$$E(X) = \frac{1}{n}\,\frac{n(n+1)}{2} = \frac{1}{2}(n+1).$$

As you might expect, this is the 'middle' of the range of X: it is midway between the lowest and highest values.

Solution to Activity 12

The formula $E(X) = (n+1)/2$ gives the following means.

(a) When $n = 4$, $\frac{1}{2}(4+1) = \frac{5}{2} = 2.5$.

(b) When $n = 9$, $\frac{1}{2}(9+1) = \frac{10}{2} = 5$.

Solution to Activity 13

The formula $E(X) = (n+m)/2$ gives the following mean when $n = 9$ and $m = 0$: $\frac{1}{2}(9+0) = \frac{9}{2} = 4.5$. The inclusion of zero has reduced the mean from 5 (for $1, 2, \ldots, 9$) to 4.5 (for $0, 1, \ldots, 9$).

Solution to Activity 14

We use Equation (12) in both parts of the activity.

(a) $$E(X) = \int_0^1 x\, f(x)\, dx = \int_0^1 x(3x^2)\, dx$$
$$= 3\int_0^1 x^3\, dx = 3\left[\frac{x^4}{4}\right]_0^1 = \frac{3}{4}(1-0) = \frac{3}{4}.$$

(b) $$E(X) = \int_1^2 x\, f(x)\, dx = \int_1^2 x(0.6x^2 + 0.2x - 0.7)\, dx$$
$$= \int_1^2 (0.6x^3 + 0.2x^2 - 0.7x)\, dx$$
$$= \left[0.15x^4 + \frac{0.2x^3}{3} - 0.35x^2\right]_1^2$$
$$\simeq 2.4 + 0.5333 - 1.4 - (0.15 + 0.0667 - 0.35) \simeq 1.667.$$

Solution to Activity 15

$$E(X) = \int_3^6 x\, f(x)\, dx = \int_3^6 x\, \frac{1}{30}(10x - x^2 - 14)\, dx$$
$$= \frac{1}{30}\int_3^6 (10x^2 - x^3 - 14x)\, dx$$
$$= \frac{1}{30}\left[\frac{10x^3}{3} - \frac{x^4}{4} - 7x^2\right]_3^6$$
$$= \frac{1}{30}\{720 - 324 - 252 - (90 - 20.25 - 63)\} = 4.575.$$

Solution to Activity 16

(a) Either by analogy with the discrete uniform distribution or by a balancing argument like that in Example 6, you might expect the mean of the continuous uniform distribution to be halfway between the upper and lower bounds of the range of the distribution.

(b) $$E(X) = \int_a^b x\, f(x)\, dx = \int_a^b \frac{x}{b-a}\, dx = \frac{1}{b-a}\int_a^b x\, dx$$
$$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}.$$

This is the value halfway between the upper and lower bounds of the range of the distribution.

Solution to Activity 17

We use the fact (from Activity 16(b)) that the mean is given by $(a+b)/2$.

(a) $E(X) = (0+40)/2 = 20.$

(b) $E(Y) = (0+1)/2 = 1/2.$

Solution to Activity 18

The probability distribution of W and the calculations required to find the variance of W are shown in Table 15. From Activity 1, the mean μ is 4.

Table 15

w	1	3	4	5	6
$w-\mu$	-3	-1	0	1	2
$(w-\mu)^2$	9	1	0	1	4
$p(w)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$

The variance of W is

$$\begin{aligned}\sigma^2 = V(W) &= \sum (w-\mu)^2 p(w)\\ &= \left(9\times\tfrac{1}{6}\right)+\left(1\times\tfrac{1}{6}\right)+\left(0\times\tfrac{1}{6}\right)+\left(1\times\tfrac{1}{3}\right)+\left(4\times\tfrac{1}{6}\right)\\ &= \frac{9+1+0+2+4}{6} = \frac{16}{6} \simeq 2.67.\end{aligned}$$

The variance of the score on an unbiased die is 2.92. So the variance of the score on this biased die is smaller than that on an unbiased die.

Solution to Activity 19

(a) Using Equation (15) with $p = \frac{1}{2}$, the variance of the score is

$$\sigma^2 = V(X) = \tfrac{1}{2}\left(1-\tfrac{1}{2}\right) = \tfrac{1}{2}\times\tfrac{1}{2} = \tfrac{1}{4}.$$

(b) Since $p = \frac{1}{3}$, the variance of X is

$$\sigma^2 = V(X) = \tfrac{1}{3}\left(1-\tfrac{1}{3}\right) = \tfrac{1}{3}\times\tfrac{2}{3} = \tfrac{2}{9}.$$

Solution to Activity 20

(a) Using Equation (17) with $n = 100$ and $p = 0.01$, the variance of the number of defective items in a sample of size 100 is

$$\sigma^2 = 100\times 0.01\times(1-0.01) = 100\times 0.01\times 0.99 = 0.99.$$

(b) The number of arrows that hit the centre of the target has a $B\left(10, \frac{3}{4}\right)$ distribution, so the variance is

$$\sigma^2 = 10\times\tfrac{3}{4}\times\left(1-\tfrac{3}{4}\right) = 10\times\tfrac{3}{4}\times\tfrac{1}{4} = \tfrac{15}{8} = 1.875.$$

(c) The number of wins has a $B(5, 0.7)$ distribution, so the variance is

$$\sigma^2 = 5\times 0.7\times(1-0.7) = 5\times 0.7\times 0.3 = 1.05.$$

Solution to Activity 21

(a) The number of times a fair coin needs to be tossed to obtain heads has a $G\left(\frac{1}{2}\right)$ distribution, so using Equation (18) with $p = \frac{1}{2}$, the variance of this number is

$$\sigma^2 = \tfrac{1}{2}\Big/\left(\tfrac{1}{2}\right)^2 = 2.$$

(b) The number of times an unbiased die needs to be rolled to obtain a six has a $G(\frac{1}{6})$ distribution, so the variance of this number is

$$\sigma^2 = \tfrac{5}{6}\big/\left(\tfrac{1}{6}\right)^2 = 5 \times 6 = 30.$$

Solution to Activity 22

Equation (19) gives the following variances.

(a) $V(X) = 0.5$. (b) $V(Y) = 0.6825$.

Solution to Activity 23

For the binomial distribution,

$$V(X) = (1-p) \times np = (1-p) \times E(X).$$

Since $0 < p < 1$, it is also the case that $0 < 1 - p < 1$. In particular, this means that $V(X)$ is always less than $E(X)$: the binomial distribution is under-dispersed for all (allowable) values of n and p.

Solution to Activity 24

(a) If X is the score obtained when an unbiased six-sided die is rolled, then X has a discrete uniform distribution with $m = 1$ and $n = 6$. So

$$V(X) = \frac{(6-1)(6-1+2)}{12} = \frac{35}{12} \simeq 2.92.$$

(This confirms the value found without using the general formula in Example 8.)

(b) Since Y has a discrete uniform distribution with parameters $m = 1$ and $n = 9$,

$$V(Y) = \frac{(9-1)(9-1+2)}{12} = \frac{20}{3} \simeq 6.67.$$

(c) Since Z has a discrete uniform distribution with parameters $m = 0$ and $n = 9$,

$$V(Z) = \frac{(9-0)(9-0+2)}{12} = \frac{33}{4} = 8.25.$$

The addition of zero to the range of Z compared with that of Y has resulted in an increase in variance.

Solution to Activity 25

(a) Here, $a = 0$, $b = 40$, so

$$V(X) = \frac{(40-0)^2}{12} = \frac{400}{3} \simeq 133.3.$$

(b) Here, $a = 0$, $b = 1$, so

$$V(Y) = \frac{(1-0)^2}{12} = \frac{1}{12} \simeq 0.083.$$

Also, $S(Y) = \sqrt{1/12} \simeq 0.289$.

(c) The standard deviation of the $U(a, b)$ distribution cannot be $(a - b)/\sqrt{12}$ because this is a negative quantity: $a < b$. The standard deviation of the $U(a, b)$ distribution is $(b - a)/\sqrt{12}$: always take the non-negative square root of the variance.

Solution to Activity 26

$$\begin{aligned}\sigma^2 &= \int_0^1 (x-\mu)^2 f(x)\,dx = \int_0^1 \left(x - \frac{3}{4}\right)^2 3x^2\,dx \\ &= 3\int_0^1 \left(x^4 - \frac{3}{2}x^3 + \frac{9}{16}x^2\right) dx \\ &= 3\left[\frac{x^5}{5} - \frac{3x^4}{8} + \frac{3x^3}{16}\right]_0^1 \\ &= 3\left(\frac{1}{5} - \frac{3}{8} + \frac{3}{16} - 0\right) = \frac{3}{80} = 0.0375.\end{aligned}$$

Solution to Activity 27

In Example 12, $a = 2.54$, $b = 0$; in Example 13, $a = 1$, $b = -1$; and in Example 14, $a = 9/5$, $b = 32$.

Solution to Activity 28

(a) Since any value of X is shifted along by b, it would seem reasonable to expect the average value, $E(X)$, to be shifted along by b also, that is, $E(Y) = E(X) + b = \mu + b$. This proves to be correct.

(b) The variability or dispersion in the values of X would seem not to be affected by adding the same constant to each value. So it would seem reasonable to expect $V(Y) = V(X) = \sigma^2$. This proves to be correct also.

Solution to Activity 29

$$E(Y) = E(X - 1) = E(X) - 1 = \frac{1}{p} - 1 = \frac{1-p}{p}.$$

Solution to Activity 30

If X, the water temperature measured in °C, has mean 26, then the random variable $Y = \frac{9}{5}X + 32$, which represents the water temperature in °F, has mean

$$E(Y) = E\left(\tfrac{9}{5}X + 32\right) = \tfrac{9}{5}\,E(X) + 32 = \left(\tfrac{9}{5} \times 26\right) + 32 = 78.8.$$

Solution to Activity 31

$$V(Y) = V(X - 1) = 1 \times V(X) = \frac{1-p}{p^2}.$$

Solution to Activity 32

If X, the water temperature measured in °C, has standard deviation 1.25, then the random variable $Y = \frac{9}{5}X + 32$, which represents the water temperature in °F, has standard deviation

$$S(Y) = S\left(\tfrac{9}{5}X + 32\right) = \tfrac{9}{5}\,S(X) = \tfrac{9}{5} \times 1.25 = 2.25$$

and variance

$$V(Y) = \{S(Y)\}^2 = (2.25)^2 = 5.0625.$$

Solution to Activity 33

(a) The mean of Y is given by

$$\begin{aligned} E(Y) &= \left(0 \times \tfrac{1}{7}\right) + \left(1 \times \tfrac{2}{7}\right) + \left(2 \times \tfrac{3}{7}\right) + \left(3 \times \tfrac{1}{7}\right) \\ &= \frac{0+2+6+3}{7} = \frac{11}{7} \simeq 1.57. \end{aligned}$$

(b) The expected value of Y^2 is

$$\begin{aligned} E(Y^2) &= \left(0^2 \times \tfrac{1}{7}\right) + \left(1^2 \times \tfrac{2}{7}\right) + \left(2^2 \times \tfrac{3}{7}\right) + \left(3^2 \times \tfrac{1}{7}\right) \\ &= \frac{0+2+12+9}{7} = \frac{23}{7}. \end{aligned}$$

So, using Equation (29), the variance of Y is

$$V(Y) = \frac{23}{7} - \left(\frac{11}{7}\right)^2 = \frac{40}{49} \simeq 0.82.$$

This is an example where using the alternative formula for the variance simplifies the calculations and avoids introducing rounding error.

Solution to Activity 34

If X follows the $U(a, b)$ distribution, then

$$\begin{aligned} E(X^2) &= \int_a^b x^2 f(x)\,dx = \int_a^b \frac{x^2}{b-a}\,dx = \frac{1}{b-a}\int_a^b x^2\,dx \\ &= \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}, \end{aligned}$$

using the hint in the question. Therefore

$$\begin{aligned} V(X) &= E(X^2) - \mu^2 \\ &= \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12} \\ &= \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}, \end{aligned}$$

as required.

Solution to Activity 35

$$
\begin{aligned}
E(X^2) &= \int_3^6 x^2 f(x)\,dx = \int_3^6 x^2 \frac{1}{30}(10x - x^2 - 14)\,dx \\
&= \frac{1}{30}\int_3^6 (10x^3 - x^4 - 14x^2)\,dx \\
&= \frac{1}{30}\left[\frac{5x^4}{2} - \frac{x^5}{5} - \frac{14x^3}{3}\right]_3^6 \\
&= \frac{1}{30}\{3240 - 1555.2 - 1008 - (202.5 - 48.6 - 126)\} \\
&= 21.63.
\end{aligned}
$$

Hence

$$V(X) = E(X^2) - \mu^2 = 21.63 - (4.575)^2 \simeq 0.699 \text{ cm}^2$$

and

$$S(X) = \sqrt{V(X)} \simeq 0.836 \text{ cm}.$$

Solutions to exercises

Solution to Exercise 1

(a) The expected value of X is

$$\begin{aligned} E(X) &= (2 \times 0.1) + (3 \times 0.2) + (4 \times 0.3) + (5 \times 0.4) \\ &= 0.2 + 0.6 + 1.2 + 2.0 = 4.0. \end{aligned}$$

(b) The mean of the random variable Y is

$$\begin{aligned} E(Y) &= (0 \times 0.4) + (1 \times 0.2) + (2 \times 0.1) + (3 \times 0.1) + (4 \times 0.2) \\ &= 0 + 0.2 + 0.2 + 0.3 + 0.8 = 1.5. \end{aligned}$$

Solution to Exercise 2

The mean number of people in a family of 6 who contract the disease is

$$\begin{aligned} \mu &= \left(1 \times \frac{3}{90}\right) + \left(2 \times \frac{8}{90}\right) + \left(3 \times \frac{15}{90}\right) + \left(4 \times \frac{20}{90}\right) + \left(5 \times \frac{24}{90}\right) + \left(6 \times \frac{20}{90}\right) \\ &= \frac{3 + 16 + 45 + 80 + 120 + 120}{90} = \frac{384}{90} \simeq 4.3. \end{aligned}$$

Solution to Exercise 3

(a) Since $X \sim$ Bernoulli(0.8), the mean of X is 0.8 (using Equation (4)).

(b) If Y is a random variable which represents the number of arrows out of 7 that hit the centre of the target, then $Y \sim B(7, 0.8)$. So using Equation (6), the expected number of arrows that hit the centre of the target is

$$E(Y) = 7 \times 0.8 = 5.6.$$

(c) The probability that each arrow misses the centre of the target is $1 - 0.8 = 0.2$. So if W is the number of arrows shot, up to and including the first that misses the centre of the target, then W has a geometric distribution with parameter $p = 0.2$: $W \sim G(0.2)$. So using Equation (7), the expected number of arrows the archer shoots is

$$E(W) = \frac{1}{0.2} = 5.$$

Solution to Exercise 4

We use Equation (12) in each part of the exercise.

(a) $$\begin{aligned} E(X) &= \int_0^1 x\, f(x)\, dx = \int_0^1 x(4x^3)\, dx \\ &= 4\int_0^1 x^4\, dx = 4\left[\frac{x^5}{5}\right]_0^1 = \frac{4}{5}(1 - 0) = \frac{4}{5}. \end{aligned}$$

(b) $$E(X) = \int_0^1 x\, f(x)\, dx = \int_0^1 x\left\{3(1-x)^2\right\} dx$$
$$= 3\int_0^1 (x - 2x^2 + x^3)\, dx$$
$$= 3\left[\frac{x^2}{2} - \frac{2x^3}{3} + \frac{x^4}{4}\right]_0^1$$
$$= 3\left(\frac{1}{2} - \frac{2}{3} + \frac{1}{4} - 0\right) = \frac{1}{4}.$$

(c) $$E(X) = \int_1^2 x\, f(x)\, dx = \int_1^2 x\left\{3(x-1)^2\right\} dx.$$

But $(x-1)^2 = (1-x)^2$, so

$$E(X) = \int_1^2 x\left\{3(1-x)^2\right\} dx,$$

and so, using the indefinite integral from part (b),

$$E(X) = 3\left[\frac{x^2}{2} - \frac{2x^3}{3} + \frac{x^4}{4}\right]_1^2$$
$$= 3\left\{2 - \frac{16}{3} + 4 - \left(\frac{1}{2} - \frac{2}{3} + \frac{1}{4}\right)\right\} = \frac{7}{4}.$$

Solution to Exercise 5

$$E(X) = \int_0^b x\, f(x)\, dx = \int_0^b x\, \frac{2(b-x)}{b^2}\, dx$$
$$= \frac{2}{b^2}\int_0^b (bx - x^2)\, dx = \frac{2}{b^2}\left[\frac{bx^2}{2} - \frac{x^3}{3}\right]_0^b$$
$$= \frac{2}{b^2}\left(\frac{b^3}{2} - \frac{b^3}{3} - 0\right) = \frac{2}{b^2} \times \frac{b^3}{6} = \frac{b}{3}.$$

Solution to Exercise 6

$$E(X) = \int_0^2 x\, f(x)\, dx = \int_0^2 x\frac{15}{16\sqrt{2}}\sqrt{x}(2-x)\, dx$$
$$= \frac{15}{16\sqrt{2}}\int_0^2 (2x^{3/2} - x^{5/2})\, dx = \frac{15}{16\sqrt{2}}\left[\frac{4x^{5/2}}{5} - \frac{2x^{7/2}}{7}\right]_0^2$$
$$= \frac{15}{16\sqrt{2}}\left(\frac{4 \times 4 \times \sqrt{2}}{5} - \frac{2 \times 8 \times \sqrt{2}}{7} - 0\right)$$
$$= \frac{15}{16\sqrt{2}} \times 16\sqrt{2} \times \left(\frac{1}{5} - \frac{1}{7}\right)$$
$$= 15 \times \frac{2}{35} = \frac{6}{7} \simeq 0.857 \text{ minutes}$$

(or about 51.4 seconds).

Solution to Exercise 7

(a) The probability distribution of X and the calculations required to find the variance of X are shown in Table 16. From the solution to Exercise 1(a), the mean μ is 4.

Table 16

x	2	3	4	5
$x-\mu$	-2	-1	0	1
$(x-\mu)^2$	4	1	0	1
$p(x)$	0.1	0.2	0.3	0.4

The variance of X is

$$\begin{aligned} V(X) &= \sum (x-\mu)^2 p(x) \\ &= (4 \times 0.1) + (1 \times 0.2) + (0 \times 0.3) + (1 \times 0.4) \\ &= 0.4 + 0.2 + 0 + 0.4 = 1. \end{aligned}$$

(b) The probability distribution of Y and the calculations required to find the variance of Y are shown in Table 17. From the solution to Exercise 1(b), the mean μ is 1.5.

Table 17

y	0	1	2	3	4
$y-\mu$	-1.5	-0.5	0.5	1.5	2.5
$(y-\mu)^2$	2.25	0.25	0.25	2.25	6.25
$p(y)$	0.4	0.2	0.1	0.1	0.2

The variance of Y is

$$\begin{aligned} V(Y) &= \sum (y-\mu)^2 p(y) \\ &= (2.25 \times 0.4) + (0.25 \times 0.2) + (0.25 \times 0.1) \\ &\quad + (2.25 \times 0.1) + (6.25 \times 0.2) \\ &= 0.9 + 0.05 + 0.025 + 0.225 + 1.25 = 2.45. \end{aligned}$$

Solution to Exercise 8

(a) Since $X \sim$ Bernoulli(0.8), using Equation (15), the variance of X is

$$V(X) = 0.8(1-0.8) = 0.16.$$

(b) If Y is a random variable which represents the number of arrows that hit the centre of the target in 7 shots, then $Y \sim B(7, 0.8)$. So, using Equation (17), the variance of the number of arrows that hit the centre of the target is

$$V(Y) = 7 \times 0.8 \times (1-0.8) = 7 \times 0.8 \times 0.2 = 1.12.$$

(c) The probability that each arrow misses the centre of the target is $1-0.8 = 0.2$. So if W is the number of arrows shot, up to and including the first that misses the centre of the target, then W has a geometric distribution with parameter $p = 0.2$, that is, $W \sim G(0.2)$.

So, using Equation (18), the variance of the number of arrows that the archer shoots is

$$V(W) = \frac{1 - 0.2}{0.2^2} = 20.$$

Solution to Exercise 9

$$\begin{aligned}\sigma^2 &= \int_0^1 (x-\mu)^2 f(x)\,dx = \int_0^1 \left(x - \frac{4}{5}\right)^2 4x^3\,dx \\ &= 4\int_0^1 \left(x^5 - \frac{8}{5}x^4 + \frac{16}{25}x^3\right) dx \\ &= 4\left[\frac{x^6}{6} - \frac{8x^5}{25} + \frac{4x^4}{25}\right]_0^1 \\ &= 4\left(\frac{1}{6} - \frac{8}{25} + \frac{4}{25} - 0\right) = \frac{2}{75} \simeq 0.027.\end{aligned}$$

Solution to Exercise 10

(a) The answer is $Y = 2X - 1$. This might be obvious to you but if not, write $Y = aX + b$ and find a and b such that $Y = 1$ when $X = 1$ (that is, $a + b = 1$) and $Y = -1$ when $X = 0$ (that is, $b = -1$); solving the equations in brackets gives $b = -1$, $a = 2$.

(b) Using Equation (24), $E(Y) = 2\,E(X) - 1 = 2p - 1$.

(c) Using Equation (25), $V(Y) = 2^2\,V(X) = 4p(1-p)$.

Solution to Exercise 11

(a) According to Equation (23), $T = \frac{9}{5}W + 32$. Solving for T gives

$$W = \frac{5}{9}(T - 32) = \frac{5}{9}T - \frac{160}{9}.$$

(b) Using Equation (24),

$$E(W) = \frac{5}{9}E(T) - \frac{160}{9} = \frac{4225 - 160}{9} = \frac{1355}{3} \simeq 451.7.$$

(c) Using Equation (25),

$$S(W) = \frac{5}{9}S(T) = \frac{200}{9} \simeq 22.2.$$

Solution to Exercise 12

$$\begin{aligned}E(X^2) &= \int_0^b x^2 f(x)\,dx = \int_0^b x^2 \frac{2(b-x)}{b^2}\,dx = \frac{2}{b^2}\int_0^b (bx^2 - x^3)\,dx \\ &= \frac{2}{b^2}\left[\frac{bx^3}{3} - \frac{x^4}{4}\right]_0^b = \frac{2}{b^2}\left(\frac{b^4}{3} - \frac{b^4}{4} - 0\right) = \frac{2}{b^2} \times \frac{b^4}{12} = \frac{b^2}{6}.\end{aligned}$$

Hence, using Equation (29),

$$V(X) = E(X^2) - \mu^2 = \frac{b^2}{6} - \left(\frac{b}{3}\right)^2 = \frac{b^2}{18}.$$

Solution to Exercise 13

$$\begin{aligned}
E(X^2) &= \int_0^2 x^2 f(x)\,dx = \int_0^2 x^2 \frac{15}{16\sqrt{2}}\sqrt{x}(2-x)\,dx \\
&= \frac{15}{16\sqrt{2}}\int_0^2 (2x^{5/2} - x^{7/2})\,dx = \frac{15}{16\sqrt{2}}\left[\frac{4x^{7/2}}{7} - \frac{2x^{9/2}}{9}\right]_0^2 \\
&= \frac{15}{16\sqrt{2}}\left(\frac{4\times 8\times\sqrt{2}}{7} - \frac{2\times 16\times\sqrt{2}}{9} - 0\right) \\
&= \frac{15}{16\sqrt{2}} \times 32\sqrt{2} \times \left(\frac{1}{7} - \frac{1}{9}\right) = 30 \times \frac{2}{63} = \frac{20}{21}.
\end{aligned}$$

Hence

$$V(X) = E(X^2) - \mu^2 = \frac{20}{21} - \left(\frac{6}{7}\right)^2 = \frac{32}{147} \simeq 0.218,$$

so the standard deviation is the square root of this, which is about 0.467 minutes (or about 28 seconds).

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 207: © Duncan Andison / www.123rf.com

Page 209: Taken from: http://globe-views.com/dreams/cat.html

Page 210: © Cathy Yeulet / www.123RF.com

Page 214: © VladimirGerasimov / istockphoto.com

Page 215: © oversnap / www.istockphoto.com

Page 216: © Oleg Dudko / www.123rf.com

Page 220: © James Gathany – CDC Public Health Image library ID 11162

Page 222: Taken from: http://www.thetimes.co.uk/tto/news/world/australia-newzealand/article4168304.ece

Page 223: © Ekawat Chaowicharat / www.123rf.com

Page 225: © Katn1999 / www.dreamstime.com

Page 226: © wirojsid / www.123rf.com

Page 230: © mbbirdy / www.istockphoto.com

Page 233: © kzenon / www.123rf.com

Page 237: © 2016 Under Armour, Inc. All Rights Reserved

Page 238: © Ged Carroll

Page 241: © James Cathany / https://commons.wikimedia.org/wiki/File:Culex_sp_larvae.png This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 244: Taken from: http://acidcow.com/fun/62662-acid-picdump-106-pics.html

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.